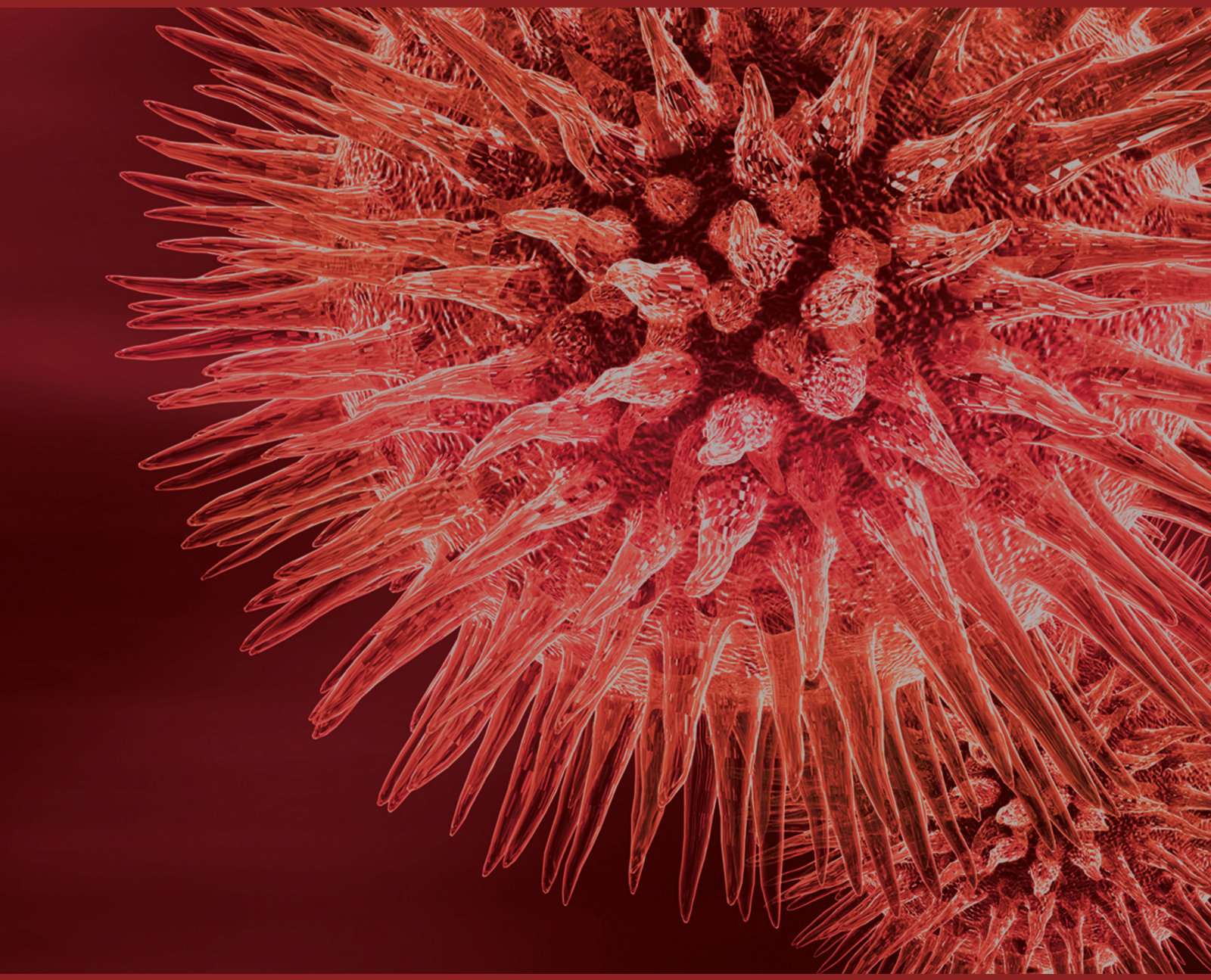
BioMed Research International

Statistical Analysis of High-Dimensional Genetic Data in Complex Traits

Guest Editors: Taesung Park, Kristel Van Steen, Xiang-Yang Lou, and Momiao Xiong



Statistical Analysis of High-Dimensional Genetic Data in Complex Traits

BioMed Research International

Statistical Analysis of High-Dimensional Genetic Data in Complex Traits

Guest Editors: Taesung Park, Kristel Van Steen, Xiang-Yang Lou, and Momiao Xiong

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 564273, 2 pages
http://dx.doi.org/10.1155/2015/564273

Editorial

Statistical Analysis of High-Dimensional Genetic Data in Complex Traits

Taesung Park,[1] Kristel Van Steen,[2] Xiang-Yang Lou,[3] and Momiao Xiong[4]

[1]*Department of Statistics, Seoul National University, Gwanak_1 Gwanak-ro, Gwanak-gu, Seoul 151-747, Republic of Korea*
[2]*Montefiore Institute, Université de Liège, Bâtiment B28, Office 0.15-B37, Grande Traverse 10, 4000 Liège, Belgium*
[3]*Department of Biostatistics, University of Alabama at Birmingham, Ryals Public Health Building, No. 420B, Birmingham, AL 35294, USA*
[4]*Division of Biostatistics, Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Herman Pressler, Houston, TX 77030, USA*

Correspondence should be addressed to Taesung Park; tspark@stats.snu.ac.kr

Received 2 June 2015; Accepted 2 June 2015

With the recent development of high-throughput DNA microarray and next-generation sequencing techniques for detecting various genomic variants (SNVs, CNVs, INDELs, etc.), genome-wide association studies (GWASs) have become a popular strategy to discover genetic factors affecting common complex diseases. Many GWASs have successfully identified genetic risk factors associated with common diseases and have achieved substantial success in unveiling genomic regions responsible for the various aspects of phenotypes.

However, identifying the underlying mechanism of disease susceptible loci has proven to be difficult due to the complex genetic architecture of common diseases. The previously associated variants through GWASs only explain a small portion of the genetic factors in complex diseases. This rather limited finding is partly ascribed to the lack of intensive analysis on undiscovered genetic determinants such as rare variants and gene-gene interactions. Unfortunately, standard methods used to test for association with single common genetic variants are underpowered for detection of rare variants and genetic interactions.

This special issue is dedicated to presenting state-of-the-art statistical and computational methods for finding missing heritability underlying complex traits with massive genetic data including GWAS, next-generation sequencing, and DNA microarray data. The main focus of this special issue is on data mining and machine learning for advanced GWAS analysis. The advanced GWAS analysis includes multi-SNP analysis, gene-gene and gene-environment interaction analysis, estimation of missing heritability, and analysis of population heterogeneity. This special issue provides a platform to the researchers with expertise in data mining to discuss recent advancements in analytic approach of post-GWAS association analysis in field of statistics and bioinformatics.

The paper by W. Lee et al. proposes an approach to identifying clinically interesting subgroups in a heterogeneous population. The identification step uses a clustering algorithm and proposes an improved false discovery rate-(FDR-) based measure to remedy the overestimation of the ordinary FDR-based approach. The paper by Y. Kim et al. performs heritability estimation by using population- and family-based samples. The main idea lies in utilizing genetic relationship matrix to parameterize the variance of a polygenic effect for population-based samples.

Three other papers consider gene-gene and gene-environment analysis. First, J. Yee et al. proposed interaction analysis for quantitative traits using entropy. Although there have been several methods proposed for gene-gene interaction using entropy, this is a robust entropy-based gene-gene interaction analysis that does not necessarily require an assumption on the distribution of trait for

quantitative traits. Second, S. Y. Lee et al. focused on identifying multi-SNP effects or gene-gene interactions for survival phenotypes. In the framework of the multifactor dimensionality reduction (MDR) method, several extensions for the survival phenotype are considered and compared to the earlier MDR method through comprehensive simulation studies. Third, the paper by H. Xu et al. proposes a new GWAS strategy for detecting gene-gene and gene-environment analysis by combining the generalized multifactor dimensionality reduction-graphics processing unit (GMDR-GPU) algorithm with mixed linear model approach. It was further employed to investigate the genetic architecture of important quality traits in rice. The reliability and efficiency of the model and analytical methods were verified through Monte Carlo simulations.

The next two papers discuss multi-SNP analysis. Y. J. Yoo et al. propose a new multi-bin linear combination (MLC) test for multiple SNP analysis. It first performs clustering analysis to find cliques, complete subnetworks of SNPs with all pairwise correlations above a threshold, and then performs MLC test. Through simulation studies, the clique-based algorithm was shown to produce smaller clusters with stronger positive correlation than other MLC tests. The paper by S. Won et al. focuses on comparing penalized and nonpenalized methods for disease prediction with large-scale genetic data. It was shown that penalized regressions are usually robust and provide better accuracy than nonpenalized methods for disease prediction.

Next, the work of J. Joo et al. considers robust genetic association tests for GWAS. How these robust tests can be applied to the replication study of GWAS and how the overall statistical significance can be evaluated using the combined test formed by p values of the discovery and replication studies were demonstrated.

Finally, the paper by L. Li and M. Xiong proposes a dynamic model for RNA-seq data analysis. To extract biologically useful transcription process from the RNA-seq data, the ordinary differential equation (ODE) model was proposed for modeling the RNA-seq data. Differential principal analysis was developed for estimation of location-varying coefficients of the ODE.

This special issue discusses the most challenging issues in multiple SNPs approaches including gene-gene interaction and introduces statistical and computational methods for data mining and machine learning for revealing hidden association network of genotype-phenotype relationship. The nine papers in this special issue provide scientists with an overview on the recent advancements in multiple SNP analysis for GWASs. We hope the papers can encourage researchers towards a more extensive use of statistical genetics and bioinformatics techniques for research in biology and medical sciences.

Acknowledgments

We thank the authors for their excellent contributions to the special issue. We also acknowledge the dedicated works of all reviewers of these papers for their critical and helpful comments.

Taesung Park
Kristel Van Steen
Xiang-Yang Lou
Momiao Xiong

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 605891, 10 pages
http://dx.doi.org/10.1155/2015/605891

Research Article

Evaluation of Penalized and Nonpenalized Methods for Disease Prediction with Large-Scale Genetic Data

Sungho Won,[1] Hosik Choi,[2] Suyeon Park,[3,4] Juyoung Lee,[4] Changyi Park,[5] and Sunghoon Kwon[6]

[1]*Department of Public Health Science, Seoul National University, Seoul, Republic of Korea*
[2]*Department of Applied Information Statistics, Kyonggi University, Suwon, Republic of Korea*
[3]*Department of Biostatistics, Soonchunhyang University, College of Medicine, Seoul, Republic of Korea*
[4]*Center for Genome Science, National Institute of Health, Osong Health Technology Administration Complex, Chungcheongbuk-do, Seoul, Republic of Korea*
[5]*Department of Statistics, University of Seoul, Seoul, Republic of Korea*
[6]*Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea*

Correspondence should be addressed to Changyi Park; cpark463@gmail.com and Sunghoon Kwon; shkwon0522@gmail.com

Received 3 December 2014; Revised 16 January 2015; Accepted 16 January 2015

Academic Editor: Momiao Xiong

Owing to recent improvement of genotyping technology, large-scale genetic data can be utilized to identify disease susceptibility loci and this successful finding has substantially improved our understanding of complex diseases. However, in spite of these successes, most of the genetic effects for many complex diseases were found to be very small, which have been a big hurdle to build disease prediction model. Recently, many statistical methods based on penalized regressions have been proposed to tackle the so-called "large P and small N" problem. Penalized regressions including least absolute selection and shrinkage operator (LASSO) and ridge regression limit the space of parameters, and this constraint enables the estimation of effects for very large number of SNPs. Various extensions have been suggested, and, in this report, we compare their accuracy by applying them to several complex diseases. Our results show that penalized regressions are usually robust and provide better accuracy than the existing methods for at least diseases under consideration.

1. Introduction

Accurate disease prediction is a central goal of clinical genetics, and much effort has been made to utilize the large-scale genetic data for a disease prediction model for complex disease. However, except for the fully penetrant Mendelian disorders, effect sizes of most disease susceptibility loci identified by genome-wide association studies (GWAS) are usually modest [1] and the presence of much larger number of genetic variants than the sample size (or so-called "large P and small N" problem) makes the construction of a disease risk prediction model intractable. For instance, the variation of predicted risk scores for each individual is proportionally related to the number of causal variants, and the accuracy of the predicted disease status decreases with the increase of the number of causal variants when the relative proportion of variance explained by causal variants is fixed [2, 3]. Also large P and small N problem prevents the estimation of the joint effect of all markers and thus prediction model building has been based only on marginal effects of variants [4–6]. Recently, various nonpenalized and penalized statistical methods have been suggested to tackle these issues. However, a comprehensive evaluation of existing methods has not been conducted yet.

The statistical methods for the disease risk prediction at the early stage were based on gene scores [4–6]. Causal variants often have additive effects on phenotypes, and a simple linear (logistic) regression can be adopted to estimate marginal effects of each variant under the assumption that there is no gene × gene and gene × environment interactions.

Then, the coded genotypes of large-scale genetic data are multiplied with their marginal effect estimates and their sum for each individual can be incorporated to build the disease risk prediction model. Multiple studies showed that gene score-based approach is practically useful for building disease risk prediction [7]. This approach is computationally very efficient, and further extensions based on best linear unbiased predictor (BLUP) have been proposed in the literature [8–10]. However, for instance, if joint effects between multiple variants are substantial or there is large linkage disequilibrium between variants, the estimated gene score can be biased and the predicted disease risk becomes less accurate.

As an alternative to gene score-based approach, one may consider statistical learning methods in disease risk prediction. Statistical learning algorithms have been successful over the past decades in various learning tasks including text categorization, fraud detection, character and image recognition, natural language processing, and marketing. Disease risk prediction can be naturally posed as a classification problem, and support vector machines (SVMs) [11] and ensemble algorithms, in particular, random forests proposed by Breiman [12], have been often shown to yield more accurate predictions than other classification algorithms [13]. In particular, SVMs have been an important tool in classification because of their accuracy and flexibility in modeling different types of data. However, these methods have some drawbacks in disease risk prediction. Effects of variants on phenotypes in the prediction models from ensemble algorithms are difficult to interpret and SVMs do not provide class conditional probabilities [14]. Therefore, in this report, we focus on the penalized methods in logistic regression.

Recently, various penalized methods have been proposed to resolve the large P and small N problem. Examples include convex penalizations such as ridge [15–17] and LASSO [18] and nonconvex penalizations such as the smoothly clipped absolute deviation (SCAD) [19] and bridge [20]. In general, penalized methods have often provided more accurate predictions and easier interpretations than nonpenalized methods, especially when the number of samples is smaller than the number of variables. Some penalized methods automatically select relevant variables by setting the estimated coefficients of irrelevant variables as exactly zero. Also penalized methods enhance the accuracy of predictions by shrinking the coefficients of nonzero elements with data-adaptive tuning parameters.

In this report, we compare the performances of various nonpenalized and penalized methods in the prediction of diseases on data from Korea Association Resource (KARE) project that is a part of Korea Genome Epidemiology Study. We select individuals with extreme phenotypes among the participants in KARE project and consider the type 2 diabetes, obesity, hypertension, and three smoking-related phenotypes. The predictive performances of those nonpenalized and penalized methods are compared by area under the curve (AUC). Our results indicate that penalized methods tend to yield more accurate predictions than nonpenalized methods although their relative performances depend on particular diseases.

2. Methods

2.1. KARE Cohort. The KARE project, with 10,038 participants living in Ansung (rural) and Ansan (urban), was initiated in 2007 for large-scale GWAS based on the Korean population. Among the participants, 10,004 individuals were genotyped for 500,568 SNPs with the Affymetrix Genome-Wide Human SNP array 5.0. We discarded SNPs with p-values for Hardy-Weinberg equilibrium (HWE) less than 10^{-6}, with genotype call rates less than 95%, or minor allele frequencies (MAF) less than 0.01, and 352,228 SNPs were left for subsequent analysis. Individuals with low call rates (<95%, $n = 401$), high heterozygosity (>30%, $n = 11$), gender inconsistencies ($n = 41$), or serious concomitant illness ($n = 101$) were excluded from analysis. We considered independent samples and excluded related or identical individuals whose computed average pairwise identical in state value was higher than that estimated from first-degree relatives of Korean sib-pair samples (>0.8, $n = 608$). In total, 8,842 individuals were analyzed. From randomly selected 20 duplicate samples, we found that genotype concordance rates exceeded 99.7%, with no single SNP excessively discordant. The population substructure was handled with EIGENSTRAT approach [21] and we chose 10 principal component scores. Missing genotypes were imputed with Beagle [22].

2.1.1. Type 2 Diabetes (T2D). T2D mainly occurs in people aged over 40, and it is diagnosed with level of glucose and hemoglobin a1c (hba1c) in blood. In our studies, individuals were selected as being affected with type 2 diabetes if their hba1c are larger than 6.5, fasting plasma glucoses are larger than or equal to 126, or 2-hour postprandial blood glucoses are larger than or equal to 200. In total, there were 1182 affected individuals, and 2364 individuals not satisfying the condition for type-2 diabetes and older than the other unaffected individuals were considered as controls. As environmental variables, we considered area (Ansan/Ansung), sex, age, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), triglyceride, and ten PC scores.

2.1.2. Obesity. Obesity status was determined by BMI. Individuals were considered as cases if their BMIs are larger than 27, and there were 1022 affected individuals in KARE cohort. We also selected 2325 individuals with BMIs less than 27 and older than the other unaffected individuals as controls. We considered area, sex, age, height, waist-hip ratio, SBP, DBP, high density lipoprotein, triglyceride, and ten PC scores as environmental variables.

2.1.3. Hypertension. Hypertension status was determined by SBP and DBP. 1035 individuals with SBPs and DBPs larger than 140 and 80, respectively, were considered as cases. 2290 individuals whose SBPs and DBPs were less than 120 and 80, respectively, were selected as controls. Environmental variables considered were area, sex, age, BMI, and ten PC scores.

2.1.4. Cigarettes Smoked per Day (CPD). For smoking-related phenotypes, we considered only male samples for predicting

smoking behaviors because the number of female smokers was very small. CPD was defined to detect the nicotine dependence of each individual. Individuals whose number of cigarettes per day was larger than 20 were defined as being addicted to nicotine, and 333 individuals were selected as cases. Individuals were chosen as controls if the number of cigarettes per day was less than 10, and 375 males were chosen as controls. Environmental variables were area, age, BMI, waist-hip ratio, triglyceride, SBP, and ten PC scores.

2.1.5. Smoking Initiation (SI). Smoking status for each individual has four categories: never smoked, former smoker, occasional smoker, and habitual smoker. Males who never smoked were defined as controls, and males who occasionally or habitually smoke were defined as cases. There were 3357 cases and 807 controls, and the same clinical variables as in CPD were environmental variables.

2.1.6. Smoking Cessation (SC). SC was defined with smoking status as SI, but we used different categories for cases and controls. Males who never smoked were defined as controls, and males who occasionally or habitually smoke or smoked before were defined as cases. The numbers of cases and controls are 2064 and 1293, respectively, and environmental variables were the same clinical variables as in CPD.

2.2. Disease Risk Prediction Model Building

2.2.1. Notations. Let y_i be a dichotomous phenotype for individual i, and affected and unaffected individuals are coded as 1 and 0, respectively. The sample size is denoted as $n = n_a + n_u$, where n_a and n_u denote the numbers of cases and controls, respectively. We assume that there are p_1 genetic variants and p_2 environmental variants including an intercept. Therefore the total number of variables is $p = p_1 + p_2$. $\mathbf{x}_i$ denotes a vector with p covariates for individual i, and the coded genotypes of the kth variant and the lth environmental variable were denoted by x_{1ik} and x_{2il}, respectively. The coefficient vector of p covariates is denoted by $\boldsymbol{\beta}$.

2.2.2. Cross Validation. To see the effect of sample size, we selected n individuals where n_a cases and n_u controls with extreme phenotypes were chosen, and the relative ratios of n_a to n_u are assumed to be equal to their ratios between all available cases and controls in KARE cohorts. We evaluated the accuracies of the disease risk prediction models for different choices of n. Accuracies of the disease prediction models were assessed via 10-fold cross validation, and AUCs were calculated with 10 replicates. All individuals were randomly divided into 10 different subgroups with the same number of cases and controls. Each subgroup was used as test set once across ten replicates and therefore there is no overlap between test set in different replicates.

2.2.3. Feature Selection and Risk Prediction. Numbers of available genetic markers seem to be related to the prediction accuracy, and different numbers of genetic variants were selected to build the disease risk prediction model. We choose the top p_1 genetic variants by the order of F-ratio from train set. If we let $\overline{x}_{1.k}^{(l)}$ be the average expression level of the kth variant for individuals with phenotype l and denote the overall mean expression level of the kth variant by $\overline{x}_{1.k}$, the F-ratio of the kth variant [23, 24] is defined as

$$F_k = \frac{\sum_{i=1}^{n} \sum_{l=0}^{1} I\left(y_i = l\right)\left(\overline{x}_{1.k}^{(l)} - \overline{x}_{1.k}\right)^2}{\sum_{i=1}^{n} \sum_{l=0}^{1} I\left(y_i = l\right)\left(x_{1ik} - \overline{x}_{1.k}^{(l)}\right)^2}. \quad (1)$$

Then we build the disease risk prediction model with those selected top p_1 genetic variants and p_2 environmental variables on train set and apply the prediction model to test set.

2.3. Nonpenalized Methods

2.3.1. Genetic Risk Scores (GRS). The marginal effects of covariates are tested with F-ratio [23, 24]. Then, the coded genotypes of significant variants at $\alpha = 0.05$ level are summed to calculate GRS, and GRS and environmental variables were incorporated into the logistic regressions as covariates to build the final disease risk model on train set. The disease risk scores are calculated for individuals on test set and its accuracy of disease risk prediction model is evaluated.

2.3.2. MultiBLUP. Polygenic effects explained by available SNPs can be modeled by the linear mixed model whose variance covariance matrix is parameterized with the genetic relationship matrix [25–27], and BLUP can be used to predict the disease risk by genetic effects. However, those approaches assume that effects of all SNPs are homogeneous in spite of their heterogeneity. For instance, it has been shown that MAFs of SNPs may reveal some information about genetic architecture [28] and random effects need to be defined for SNPs with different spectra of MAFs separately. MultiBLUP [10] categorizes each SNP into different classes with distinct effect sizes or linkage disequilibrium block and applies a linear mixed model with multiple random effects to improve the accuracy of the prediction model [10].

2.4. Penalized Methods. Various penalized methods have been recently proposed, and we consider five penalized methods in our comparison: ridge [29], LASSO [30], elastic-net [31], SCAD [32], and truncated ridge (TR) [33–35]. The p dimensional coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ can be estimated by minimizing the penalized negative log-likelihood:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{-y_i \mathbf{x}_i^t \boldsymbol{\beta} + \log\left(1 + \exp\left(\mathbf{x}_i^t \boldsymbol{\beta}\right)\right)\right\} + \sum_{j=1}^{p} J_\lambda\left(\left|\beta_j\right|\right), \quad (2)$$

where J_λ is a penalty function and λ is a vector of tuning parameter that can be determined by a search on an appropriate grid. Each penalized regression requires the estimation of λ, and 100 grid points of λ were considered from "glmnet" function in R for all the methods.

2.4.1. Ridge. In linear regression, estimates from least square method are quite unstable under severe multicollinearity because of their large variances. Ridge penalty

$$J_\lambda(t) = \lambda t^2 \quad (3)$$

was originally developed to stabilize the sample performance of least square estimates by shrinking their absolute values toward zero [29]. Ridge penalty controls the amount of shrinkage effect by choosing the tuning parameter λ, and the resulting ridge estimates tend to have a smaller variance than least square estimates. In particular, ridge regression can be conducted even when p is much larger than n, where the least square method does not have a model identifiability. However, ridge estimates have a drawback in the interpretation of the final model because all the covariates are included in the final model regardless of the choice of λ. Hence, ridge regression must be conducted together with an extra selection process such as stepwise subset selection or truncation methods.

2.4.2. LASSO. LASSO was proposed by Tibshirani [30] to achieve both shrinkage and covariate selection via the penalty

$$J_\lambda(t) = \lambda t. \quad (4)$$

LASSO selects relevant covariates and estimates their coefficients simultaneously by controlling the tuning parameter λ [30]. LASSO often shows a quite stable performance, especially when the sample size is small [32, 36], and achieves higher prediction accuracy than other penalized methods. LASSO has been applied to various statistical models such as Gaussian graphical models [37] because there are fast and efficient algorithms that are easily implementable [37–40]. However, several defects of LASSO have been reported in the literature [36, 41–43]. For example, LASSO tends to overfit, that is, selecting more covariates than expected [39], and is known to have a confliction between correct selection and optimal prediction [38]. To remedy such defects, modified versions of LASSO [36] were proposed and extended to the large P and small N problem [44].

2.4.3. Elastic-Net. Elastic-net penalty proposed by Zou and Hastie [31] is a convex combination of LASSO and ridge penalty is

$$J_\lambda(t) = \lambda\left(at + (1-a)t^2\right). \quad (5)$$

Here we considered 20 equally spaced grid points from zero to one for a. Elastic-net has more desirable properties than LASSO and ridge. For instance, ridge tends to keep all the covariates in the final model and hence is undesirable when there are many noncausal variants. In contrast, LASSO cannot select larger number of covariates than the sample size and tends to select a single covariate among highly correlated covariates. However, by choosing appropriate λ and a, elastic-net enables us to have balanced estimates, producing a slightly more complex model than LASSO but far simpler model than ridge. Also it achieves a grouping effect [30] on highly correlated covariates. However, elastic-net shares the disadvantage of LASSO; that is, it often overfits, which can be resolved by applying a data adaptive weight vector [45].

2.4.4. SCAD. The SCAD penalty introduced by Fan and Li [32] is

$$\frac{\partial J_\lambda(t)}{\partial t} = \min\left\{\lambda, \frac{(a\lambda - t)_+}{a-1}\right\} \quad \text{for some } a > 2, \quad (6)$$

and $a = 50$ is used for our own optimization algorithm. SCAD has several desirable properties over LASSO [32]. First, SCAD produces the same unbiased estimates as usual nonpenalized estimates of the covariates selected by SCAD. Hence SCAD can be considered as a stable version of best subset selection [46], achieving a unique benefit of the unbiased coefficient estimate [32]. Second, SCAD is known to have the oracle property [32]; that is, the set of selected covariates are asymptotically equal to the set of true causal variants. However, in spite of theoretical optimality of SCAD [47], its estimates can be poor unless the sample size is large and the effects of signal covariates are strong. In addition, similarity between numerically estimated values and theoretical ones cannot be measured because of the nonconvexity of SCAD penalty, and the computational cost for SCAD is often much more expensive than LASSO.

2.4.5. TR. As we mentioned above, ridge cannot be directly used in identifying important covariates. However, TR [35] can produce sparse estimates and inherits the same shrinkage effect as ridge that results in high prediction accuracy in the presence of multicollinearity [48]. To obtain TR estimates, we first obtain usual ridge estimates with tuning parameter λ and then truncate them with truncating level a. Hence TR declares the ridge coefficients whose absolute values smaller than a as zero and keeps the other large coefficients intact. An appropriate choice of truncating level enables us to identify a correct model while the final estimates still keep the same shrinkage property as ridge [33–35], and 20 grid points equally spaced in logarithmic scale from 0.01 to 0.001 were considered for a.

3. Results

To see the differences of penalized methods, we calculated AUCs of those methods on test set and the number of nonzero coefficients as a function of sample size. Figure 1 shows that relative performance of each method substantially depends on phenotypes, and least AUCs are often observed for SI, followed by SC. Their least AUCs may be explained by the relative importance of genetic components for each phenotype. We calculated the relative proportion of variances, h^2, explained by genotyped variants with GCTA program [27, 49]. h^2 for binary traits was estimated with all available samples by using default options, and Table 1 shows estimates for h^2. In particular, the proportion between cases and controls for each phenotype is different from true prevalence, and the ascertainment bias often happens. However the performance of each method may be related to unadjusted estimates of h^2 and ascertainment bias was not taken into account. According to Table 1, the genotyped variants explain around 25% of phenotypic variances for hypertension and CPD. However the standard error of h^2

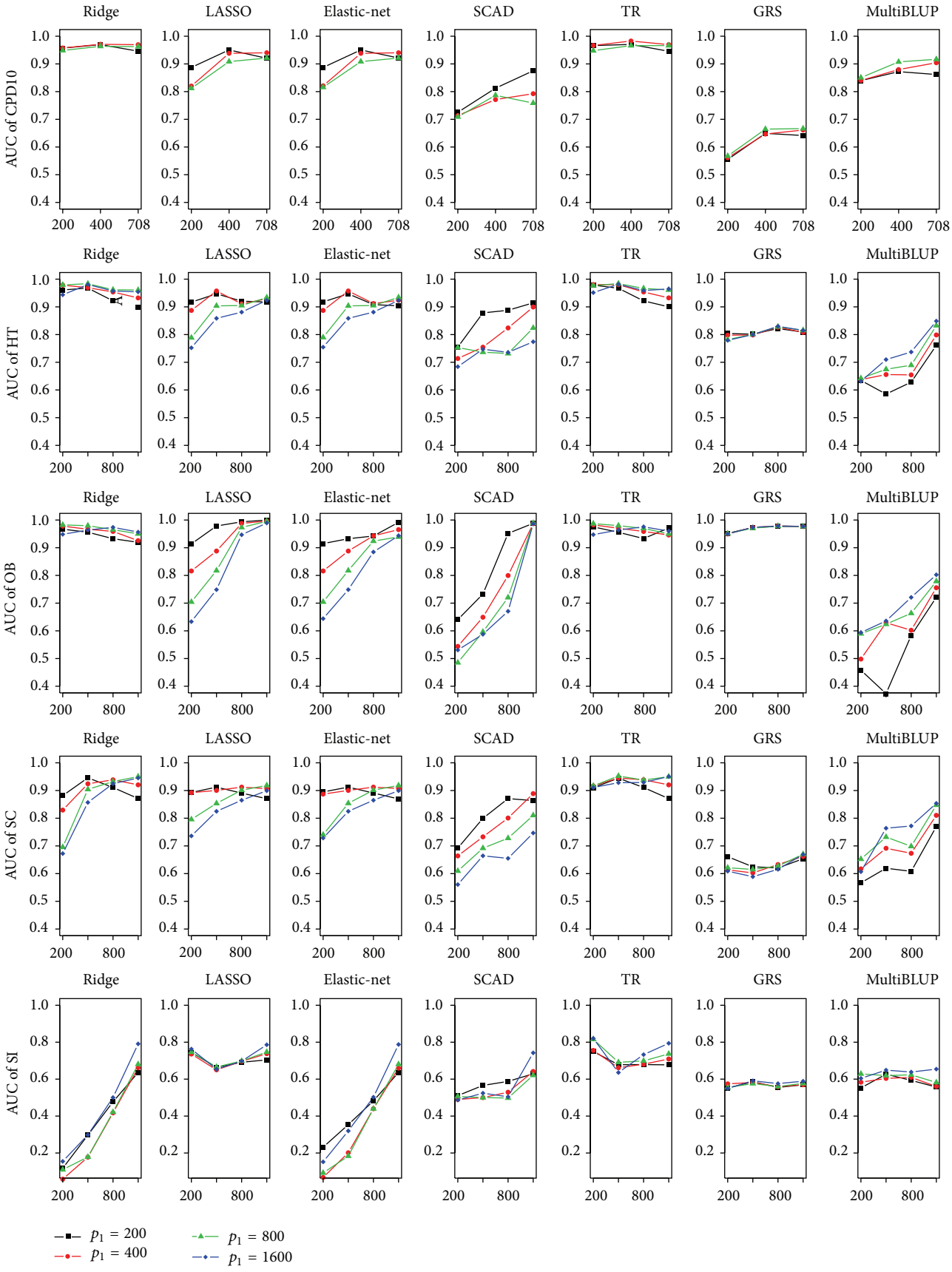


Figure 1: Continued.

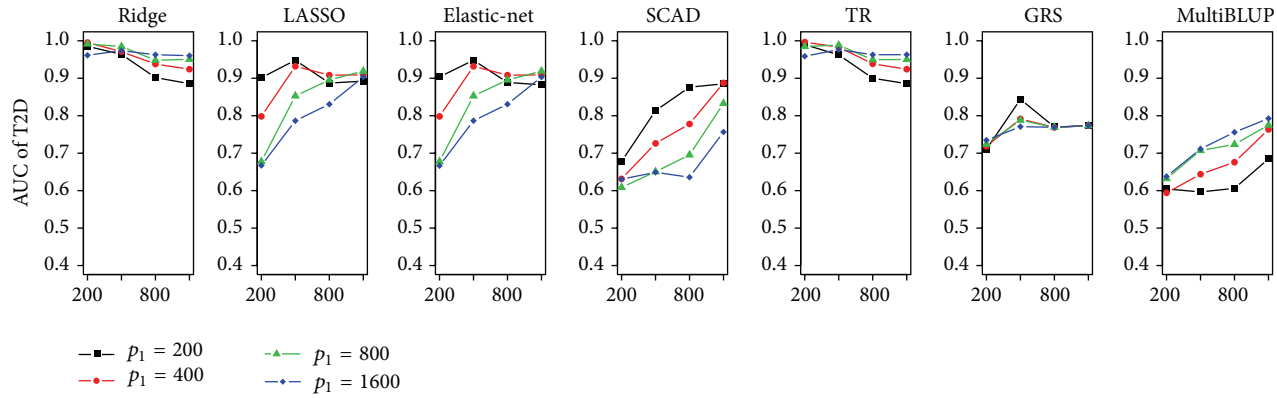


Figure 1: AUCs from test set. AUCs for T2D, obesity, hypertension, CPD10, SC, and SI from test set were calculated for different n and p_1. TR indicates the truncated ridge.

Table 1: Relative proportion of variance explained by genotyped SNPs.

	T2D	Obesity	Hypertension	CPD	SI	SC
h^2	0.147276	0.14922	0.296246	0.243554	0.052088	$1.00E-06$
$\sigma(h^2)$	0.097091	0.10029	0.100675	0.424123	0.080256	0.102595

for CPD10 is large, and genetic components for all smoking-related phenotypes seem relatively less informative.

Figure 1 shows that AUCs of two nonpenalized methods, GRS and MultiBLUP, on test set were generally outperformed by the penalized methods across various levels of n and p_1. Both approaches do not consider the joint effects among multiple causal SNPs. GRS assumes that effect sizes for causal SNPs are homogeneous and MultiBLUP assumes that sums of each SNP affect the normal distribution. However, penalized methods estimate individual effects of each SNP by shrinking each coefficient. This may explain the superiority of penalized methods over nonpenalized methods, but if those assumptions for nonpenalized methods are satisfied, they may perform better than penalized methods approaches. Interestingly MultiBLUP performs better than GRS except OB if n is larger, and AUC improvement of MultiBLUP for larger n is more substantial than GRS. Therefore, MultiBLUP seems to be more reasonable choice than GRS. Comparing overall performances of penalized methods, it can be seen that ridge and TR are the best, LASSO and elastic-net are the second, and SCAD is the last even though the performance of each method depends on specific diseases and the levels of n and p_1. Regardless of n and p_1, ridge was the best performer even for small n for all phenotypes except SI. For SI, it seems that the performance of ridge depends on n rather than p_1. TR virtually has almost the same prediction accuracy, and Figure 2 shows that its model complexity is similar with ridge for CPD, obesity, hypertension, and T2D. This observation is also strengthened by the fact that the optimal value of truncation parameter, a, is around 0.001, and thus the effect of truncation parameter on model complexity is almost negligible for these data sets. However, Figure 1 shows that differences between ridge and TR are substantial for SC and SI. AUCs of TR depend on p_1 and, in particular, are large even when n is small, which indicates that AUCs of TR depend less on n than ridge. Robustness of TR can be partially explained by smaller model complexity than ridge in Figure 2. For instance, TR usually selects quite small number of SNPs (at most 15.3 SNPs for $n \leq 800$ and 46.6 SNPs for $p_1 \leq 800$) but achieves higher prediction accuracy than ridge when n is less than 800. However, when $p_1 = n = 1600$, TR selects the same number of SNPs as ridge. Thus, we can conclude that the effect of truncation parameter diminishes for large n, which explains higher prediction accuracy when n is small.

LASSO and elastic-net show relatively large dependency on n and p_1 in prediction accuracy and model complexity for whole phenotypes except SI, and their AUCs are proportionally related to n but inversely related to p_1. Although their prediction accuracies are lower than those of ridge and TR for small n, they perform as well as ridge with small numbers of SNPs for large n. For instance, LASSO includes about 100 SNPs for $n = 1600$ and $p_1 = 200$ and about 500 SNPs for $n = 1600$ and $p_1 = 1600$, which indicates that we can construct prediction models without using the whole SNPs. Elastic-net tends to behave quite similarly as LASSO and it selects slightly larger number of SNPs for whole phenotypes except SI.

In terms of model complexity, there are substantial differences among penalized methods. Figure 2 show that SCAD selects the smallest number of covariates, while other methods such as LASSO and elastic-net usually include much more covariates. Ridge always includes all covariates, and model complexity for TR depends on data. However, even though SCAD generates the simplest model, SCAD is less preferable if it achieves the least performance among penalized regressions. For $n = 1600$, SCAD performs as well as other methods while still keeping small number of SNPs. For instance, for obesity, AUCs of SCAD are virtually the best and select extremely sparse models that have only 7.3 and 3.4 SNPs for $p_1 = 200$ and $p_1 = 1600$, respectively. Therefore, we

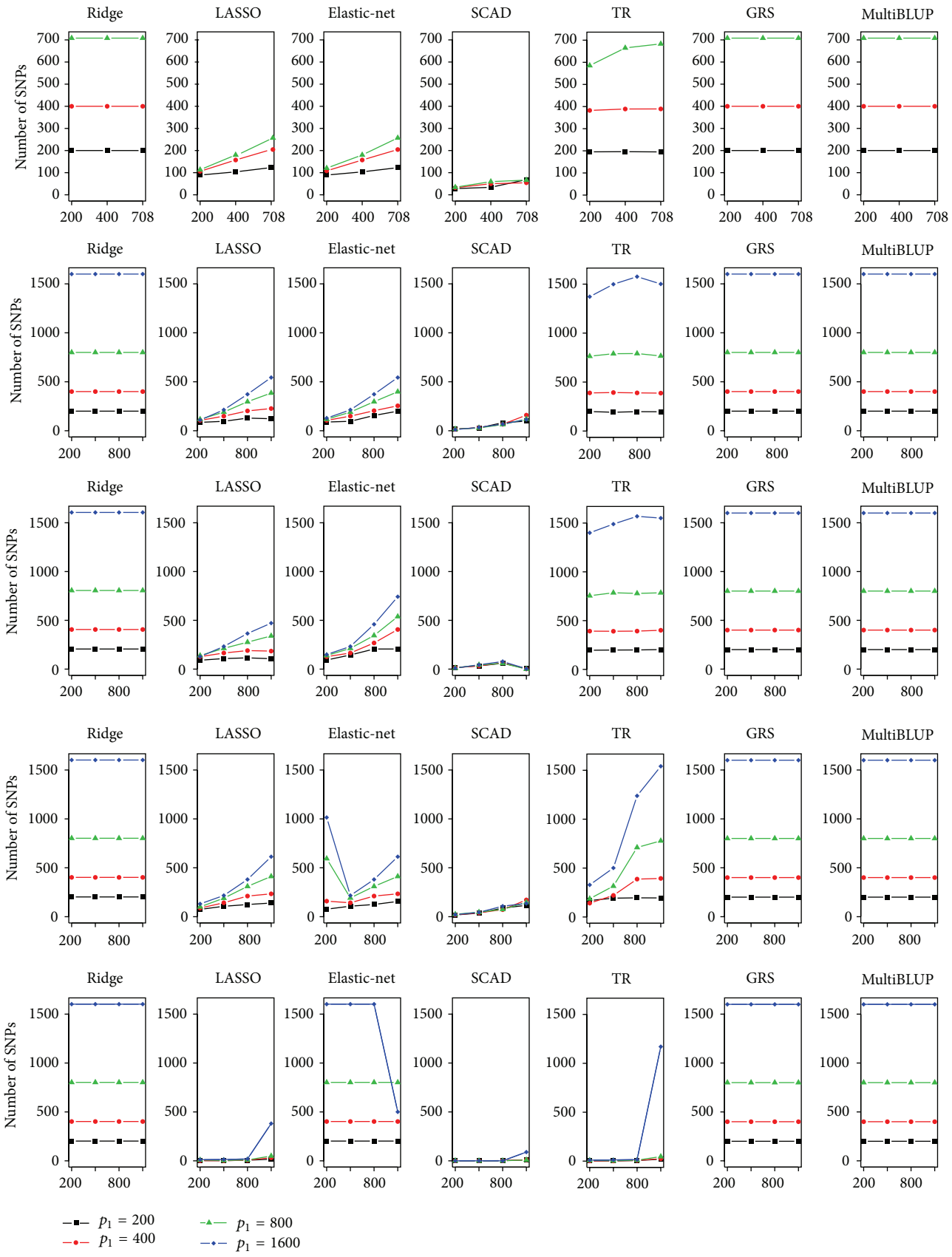


Figure 2: Continued.

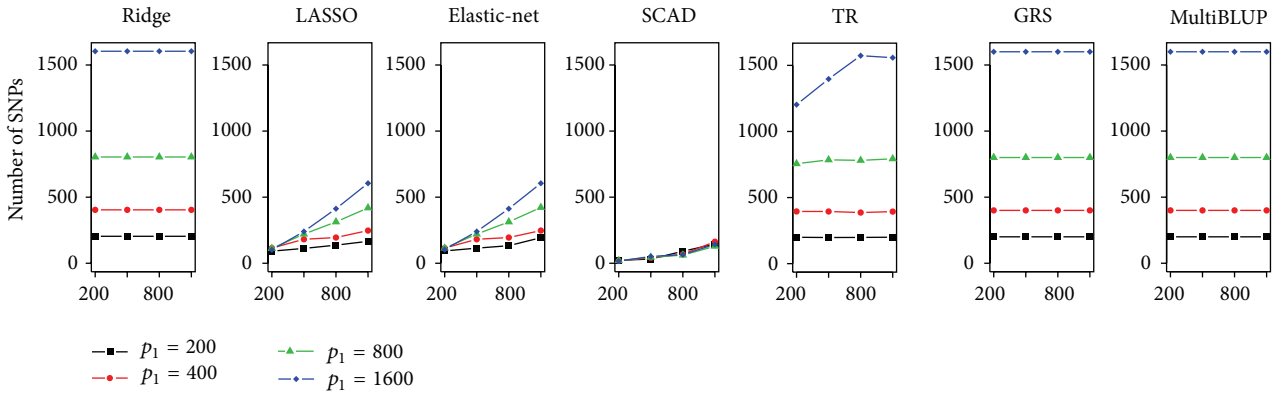


FIGURE 2: Number of nonzero p_1 in the disease risk prediction model. Numbers of nonzero coefficients of SNPs in disease risk prediction model were provided for different n and p_1. TR indicates the truncated ridge.

can conclude that SCAD is appropriate as long as relatively large number of individuals is available.

4. Discussion

In this study we have considered five penalized and two non-penalized statistical methods with six case-control datasets that are computationally feasible at the genome-wide scale. Each method was utilized to build the disease risk prediction model with different sample sizes and numbers of variants, and the accuracy of disease risk prediction models was evaluated with cross validation. Cross validation tends to overestimate the prediction accuracy, and results should be interpreted with care. A more reliable but time-consuming way is to compare the methods on random partitions of data. However cross validation does not have a strong preference towards a specific method and it may give us a rough idea on prediction accuracies of methods. According to our results, dense methods such as ridge and TR are usually more accurate than sparse methods such as LASSO and SCAD. For a large sample size, prediction accuracies from penalized methods are expected to be similar to that from ridge [23, 35, 50].

However, in spite of our comprehensive evaluations, various factors such as filtering conditions for SNPs or individuals, test statistic for prescreening, and ways of obtaining tuning parameters can affect the accuracy of the final risk prediction model, and depending on their choices, accuracy of disease risk prediction model can be substantially different. In this context, the 1-standard deviation rule [14] for tuning parameters was adopted to reduce overfitting problem. However, it did not provide any significant improvement in the results, which may indicate that there may be many causal genetic variants with small effects in the analyzed data sets. This consistently explains the reason why dense methods outperformed sparse methods such as LASSO and SCAD in our analysis. Moreover, while the results from SCAD were quite unstable for $a = 10$, the choice of $a = 50$ led to the better prediction accuracy. These findings suggest that most of SNPs have a small causal effect on diseases considered in this report. In this sense, sparse methods such as SCAD may not be preferred for infinitesimal model [51] unless the sample size is sufficiently large.

In this report we have compared various penalized regression methods. However, we have not considered more recent methods such as bootstrapping methods [33, 52, 53]. Most of them usually suffer from intensive computational burden induced by tuning extra parameters such as bootstrap size, and thus they are not computationally feasible at genome-wide scale. Alternatively, in the follow-up studies, we pursue the direction of refining the penalized methods considered in this report because there is still a significant room for improvement.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Sungho Won and Hosik Choi contributed equally to this work.

Acknowledgments

The research of S. Won was supported by the Industrial Core Technology Development Program (10040176, Development of Various Bioinformatics Software Using Next Generation Bio-Data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) and by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028559). The research of H. Choi was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (no. 2013R1A1A2007611). The research of J. Lee was provided by an intramural grant from the Korean National Institute of Health (2013-NG73002-00). The research of C. Park was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (no. 2012R1A1A2004901).

The research of S. Kwon was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002995).

References

[1] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[2] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genetics*, vol. 9, no. 3, Article ID e1003348, 2013.

[3] S. H. Lee and N. R. Wray, "Novel genetic analysis for case-control genome-wide association studies: quantification of power and genomic prediction accuracy," *PLoS ONE*, vol. 8, no. 8, Article ID e71494, 2013.

[4] D. M. Evans, P. M. Visscher, and N. R. Wray, "Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk," *Human Molecular Genetics*, vol. 18, no. 18, pp. 3525–3531, 2009.

[5] S. M. Purcell, N. R. Wray, J. L. Stone et al., "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, no. 7256, pp. 748–752, 2009.

[6] M. J. Machiela, C.-Y. Chen, C. Chen, S. J. Chanock, D. J. Hunter, and P. Kraft, "Evaluation of polygenic risk scores for predicting breast and prostate cancer risk," *Genetic Epidemiology*, vol. 35, no. 6, pp. 506–514, 2011.

[7] A. C. J. W. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects," *Human Molecular Genetics*, vol. 17, no. 2, pp. R166–R173, 2008.

[8] X. Zhou, P. Carbonetto, and M. Stephens, "Polygenic modeling with bayesian sparse linear mixed models," *PLoS Genetics*, vol. 9, no. 2, Article ID e1003264, 2013.

[9] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome Research*, vol. 17, no. 10, pp. 1520–1528, 2007.

[10] D. Speed and D. J. Balding, "MultiBLUP: improved SNP-based prediction for complex traits," *Genome Research*, vol. 24, no. 9, pp. 1550–1557, 2014.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] D. Yoon, Y. J. Kim, and T. Park, "Phenotype prediction from genome-wide association studies: application to smoking behaviors," *BMC Systems Biology*, vol. 6, no. 2, article S11, 2012.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, NY, USA, 2009.

[15] A. E. Hoerl, "Ridge regression," *Biometrics*, vol. 26, article 603, 1970.

[16] A. E. Hoerl and R. W. Kennard, "Ridge regression: applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.

[17] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[19] J. Q. Fan and R. Z. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[20] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *Journal of Machine Learning Research*, vol. 11, pp. 1081–1107, 2010.

[21] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[22] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," *The American Journal of Human Genetics*, vol. 81, no. 5, pp. 1084–1097, 2007.

[23] Y. Kim, S. Kwon, and S. H. Song, "Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data," *Computational Statistics & Data Analysis*, vol. 51, no. 3, pp. 1643–1655, 2006.

[24] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[25] S. H. Lee, J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray, "Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood," *Bioinformatics*, vol. 28, no. 19, pp. 2540–2542, 2012.

[26] S. H. Lee, D. Harold, D. R. Nyholt et al., "Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis," *Human Molecular Genetics*, vol. 22, no. 4, pp. 832–841, 2013.

[27] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *The American Journal of Human Genetics*, vol. 88, no. 1, pp. 76–82, 2011.

[28] S. H. Lee, J. Yang, G.-B. Chen et al., "Estimation of SNP heritability from dense genotype data," *The American Journal of Human Genetics*, vol. 93, no. 6, pp. 1151–1155, 2013.

[29] A. E. Horel and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[30] R. J. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996.

[31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[32] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[33] A. Chatterjee and S. N. Lahiri, "Bootstrapping lasso estimators," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 608–625, 2011.

[34] S. Kwon, Y. Kim, and H. Choi, "Sparse bridge estimation with a diverging number of parameters," *Statistics and Its Interface*, vol. 6, no. 2, pp. 231–242, 2013.

[35] J. Shao and D. Xinwei, "Estimation in high-dimensional linear models with deterministic design matrices," *The Annals of Statistics*, vol. 40, no. 2, pp. 812–831, 2012.

[36] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[37] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[38] B. Efron, T. Hastie, I. Johnstone et al., "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[39] M. Y. Park and T. Hastie, "L_1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 69, no. 4, pp. 659–677, 2007.

[40] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," *The Annals of Statistics*, vol. 35, no. 3, pp. 1012–1030, 2007.

[41] C. Leng, Y. Lin, and G. Wahba, "A note on the lasso and related procedures in model selection," *Statistica Sinica*, vol. 16, no. 4, pp. 1273–1284, 2006.

[42] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[43] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[44] J. Huang, S. Ma, and C.-H. Zhang, "Adaptive Lasso for sparse high-dimensional regression models," *Statistica Sinica*, vol. 18, no. 4, pp. 1603–1618, 2008.

[45] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *The Annals of Statistics*, vol. 37, no. 4, pp. 1733–1751, 2009.

[46] L. Breiman, "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[47] L. Wang, Y. Kim, and R. Li, "Calibrating nonconvex penalized regression in ultra-high dimension," *The Annals of Statistics*, vol. 41, no. 5, pp. 2505–2536, 2013.

[48] Z. Zheng, Y. Fan, and J. Lv, "High dimensional thresholded regression and shrinkage effect," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 76, no. 3, pp. 627–649, 2014.

[49] S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher, "Estimating missing heritability for disease from genome-wide association studies," *The American Journal of Human Genetics*, vol. 88, no. 3, pp. 294–305, 2011.

[50] C.-H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[51] G. Gibson, "Rare and common variants: twenty arguments," *Nature Reviews Genetics*, vol. 13, no. 2, pp. 135–145, 2012.

[52] P. Hall, E. R. Lee, and B. U. Park, "Bootstrap-based penalty choice for the LASSO, achieving oracle performance," *Statistica Sinica*, vol. 19, no. 2, pp. 449–471, 2009.

[53] S. Wang, B. Nan, S. Rosset, and J. Zhu, "Random lasso," *The Annals of Applied Statistics*, vol. 5, no. 1, pp. 468–485, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 135782, 7 pages
http://dx.doi.org/10.1155/2015/135782

Research Article

Detection of Epistatic and Gene-Environment Interactions Underlying Three Quality Traits in Rice Using High-Throughput Genome-Wide Data

Haiming Xu,[1] Beibei Jiang,[1] Yujie Cao,[1] Yingxin Zhang,[2] Xiaodeng Zhan,[2] Xihong Shen,[2] Shihua Cheng,[2] Xiangyang Lou,[3] and Liyong Cao[2]

[1]*Institute of Crop Science and Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China*
[2]*State Key Laboratory of Rice Biology and Zhejiang Key Laboratory of Super Rice Research, China National Rice Research Institute, Hangzhou 311401, China*
[3]*Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA*

Correspondence should be addressed to Xiangyang Lou; xylou@uab.edu and Liyong Cao; caolycgf@mail.hz.zj.cn

Received 3 December 2014; Revised 20 March 2015; Accepted 24 March 2015

Academic Editor: Ravindra N. Chibbar

With development of sequencing technology, dense single nucleotide polymorphisms (SNPs) have been available, enabling uncovering genetic architecture of complex traits by genome-wide association study (GWAS). However, the current GWAS strategy usually ignores epistatic and gene-environment interactions due to absence of appropriate methodology and heavy computational burden. This study proposed a new GWAS strategy by combining the graphics processing unit- (GPU-) based generalized multifactor dimensionality reduction (GMDR) algorithm with mixed linear model approach. The reliability and efficiency of the analytical methods were verified through Monte Carlo simulations, suggesting that a population size of nearly 150 recombinant inbred lines (RILs) had a reasonable resolution for the scenarios considered. Further, a GWAS was conducted with the above two-step strategy to investigate the additive, epistatic, and gene-environment associations between 701,867 SNPs and three important quality traits, gelatinization temperature, amylose content, and gel consistency, in a RIL population with 138 individuals derived from super-hybrid rice Xieyou9308 in two environments. Four significant SNPs were identified with additive, epistatic, and gene-environment interaction effects. Our study showed that the mixed linear model approach combining with the GPU-based GMDR algorithm is a feasible strategy for implementing GWAS to uncover genetic architecture of crop complex traits.

1. Introduction

Rice (*Oryza sativa* L.), a crop species of economic importance, provides the staple food for more than half of the population in the world. In China, the super-hybrid rice plays a pivotal role in the country's food security. There are almost eighty super-rice varieties, such as Xieyou9308, that have been successfully bred and commercially released to rice farmers since the super-rice breeding program was initiated by the Chinese government in 1996 [1]. Substantial geneticist's and breeder's effort is being expended in attempt to further investigate the mechanisms underlying high yield potential, wide adaptability, better grain quality, better disease resistance, and strong resistance to lodging in super-hybrid rice. The majority of these traits are quantitatively inherited. In addition to the increase of grain yield and the improvement of living conditions, more attention has been being paid toward improving grain quality, related to preference of cooking and eating quality of rice varieties.

As an important grain quality trait for rice, the level of amylose content (AC) is positively correlated with resistant starch (RS) content of granular starches [2–10], which is defined as the portion of dietary starch that is not digested in the small intestine of a healthy human [11]. Gelatinization

temperature (GT), the critical temperature at which about 90% of the starch granules have swelled irreversibly in hot water and start to lose crystallinity and birefringence, is another important criterion for rice quality related to cooking quality [12]. Gel consistency (GC), a measure of cold paste-viscosity of cooked milled rice flour, is a good index of cooked rice texture, especially for rice with high AC. Breeders are trying to develop high-yielding varieties with soft GC [12] because rices with soft GC cook tender and remain soft even upon cooling [13]. Therefore, understanding the genetic basis of these key traits associated with grain quality is essential to predictive rice improvement.

As high-throughput technologies producing dense single nucleotide polymorphisms (SNPs) across the whole genome, the genome-wide association study (GWAS) provides us with insightful information into genetic architecture of complex traits and is a common approach to uncover genetic components of agronomic traits. Association mapping is a high-resolution method to map quantitative trait SNPs (QTSs) based on linkage disequilibrium (LD). Association analytical methods can evaluate whether certain alleles within a population are correlated with the phenotypes of interest more frequently than the expected ones under the null hypothesis. Thus, the limitations in the traditional linkage mapping due to the statistical ambiguity with insufficient molecular markers can be alleviated. It has been widely applied in plant resource populations such as rice, maize, barley, and wheat recently [14–19].

However, the current GWAS analysis fails to detect epistatic and gene-environment interactions in most studies such as maize. But phenotypes of all living organisms represent the consequence of several genetic components including epistatic effects and their interactions with environment; therefore, to estimate genetic merit relevant to the epistases and their interactions with environment certainly plays a crucial role in planning an effective breeding regime [20, 21]. Searching for only main effects may miss the key genetic variants with specific environment response in the context of complex traits and it is not likely to provide reliable estimates of genetic component effects [22].

On the other hand, due to the prohibitively intensive computation required for a GWAS, the available methods are unpractical and difficult to extend for detection of gene-gene, gene-environment, and gene-gene-environment interactions in an experimental data with multiple environments with enormous SNPs. Currently, a workable solution is provided with the availability of generalized multifactor dimensionality reduction (GMDR) algorithm on a computing system with graphics processing units (GPUs), a type of hardware implementation of parallel computation that can be adapted for many scientific tasks [23]. The present study first used the GMDR-GPU software to screen potential candidate variants and then used the mixed liner model to dissect the epistatic and gene-environmental interactions of GT, AC, and GC in a super-hybrid rice Xieyou9308 derived RIL population, which was preferentially selected as the cardinal population for the development of BCF1 populations or immortalized F2 populations for the identification of QTLs associated with important agronomic traits [24–26]. Before analyzing the real data, Monte Carlo simulations were carried out to test the reliability and efficiency of the model and methods.

2. Materials and Methods

2.1. Field Experiment and DNA Resequencing for Genotyping. A RIL mapping population consisting of 138 lines, derived from super-hybrid rice Xieyou9308, was planted in Linshui City, Hainan Province, and Hangzhou City, Zhejiang Province, in 2009, respectively. Three quality traits, GT, AC, and GC, were investigated.

DNA resequencing was conducted in Beijing Genomics Institute (BGI) for two parents with 10X coverage and 138 lines with 2X coverage. The latest version of Nipponbare sequence [27] was used as the reference genome. Sequence alignment was conducted between the sequencing sequence and the reference genome with the software of Burrows-Wheeler Aligner (BWA) [28]. SNPs were searched between the individuals and the reference genome using the software of Sequence Alignment/Map Tools (SAMtools) [29] with the criteria of base quality over 30, mapping quality over 20, and the maximum sequence depth less than 1000. All the results were integrated by Perl program. A total of 701,867 SNPs were generated from DNA resequencing for the subsequent association study.

2.2. Genetic Models and Statistical Methods. Association mapping was performed using the mixed linear model approach. Suppose that the genetic variation of one quantitative trait is controlled by s genes. An experiment under multiple environments is conducted for gene mapping. The phenotypic value of the kth individual in the hth environment (y_{hk}) can be expressed by the following mixed linear model:

$$y_{hk} = \mu + \sum_{i=1}^{s} x_{ik} a_i + \sum_{i<j} x_{ik} x_{jk} aa_{ij} + e_h + \sum_{i} x_{ik} ae_{hi} + \sum_{i<j} x_{ik} x_{jk} aae_{hij} + \varepsilon_{hk}, \quad (1)$$

where μ is the population mean; a_i is the additive effect of the ith SNP with coefficient x_{ik}, fixed effect; aa_{ij} is the additive × additive interaction effect between the ith SNP and the jth SNP with coefficient $x_{ik}x_{jk}$, fixed effect; e_h is the random effect of the hth environment; ae_{hi} is the additive × environment interaction effect of the ith SNP and the hth environment effect with coefficient x_{ik}, random effect; aae_{hij} is the interaction effect of aa_{ij} with the hth environment with coefficient $x_{ik}x_{jk}$; and ε_{hk} is the random residual effect of the kth individual in the hth environment. The coefficient x_{ik} can be determined according to the genotype of the SNP, taking values of 1 and −1 for the homozygotes of high and low frequency alleles, respectively, and of 0 for the heterozygote.

A two-step mapping strategy for GWAS was employed to dissect genetic architecture of AC, GC, and GT. First, we used GMDR method [30] to screen SNPs potentially associated with phenotype using 1-locus model, 2-locus model, and 3-locus model, respectively; a set of reduced number of candidate SNPs was obtained for the AC, GC, and GT.

Table 1: Estimates of SNP additive effects and interaction effects with the environments under three different population structures (70%).

Chr	SNP ID	Pop	a		ae_1		ae_2		Power
			Par.	Est.	Par.	Est.	Par.	Est.	
1	28	100	−3.24	−3.32	2.65	2.67	−2.65	−2.46	99.00
		150		−3.39		2.81		−2.51	100.00
		200		−3.42		2.81		−2.45	100.00
2	100	100	−2.65	−2.79	4.05	4.16	−4.05	−3.79	100.00
		150		−2.90		4.26		−3.75	100.00
		200		−2.94		4.34		−3.73	100.00
3	93	100	−1.77	−1.86	3.24	3.34	−3.24	−3.17	87.00
		150		−1.84		3.34		−3.14	98.50
		200		−1.90		3.34		−3.09	100.00

Chr: the ordinal number for simulated chromosome; Pop: the population size; Power: the percentage of the detected SNP with significant effect at 0.05 levels; Par.: the true value of parameter in simulations; Est.: the estimate of parameter; a: additive effect; ae_1 and ae_2: the interaction effect of additive with environment 1 and 2, respectively.

Table 2: Estimates of epistasis and interaction effects with the environments under three different population structures (70%).

Chr.i	SNPID.i	Chr.j	SNPID.j	Pop	aa		aae_1		aae_2		Power
					Par.	Est.	Par.	Est.	Par.	Est.	
2	44	3	63	100	3.86	3.42	4.47	4.87	−4.47	−3.99	100.00
				150		3.24		5.05		−3.90	100.00
				200		3.21		5.12		−3.79	99.50

aa: additive-additive epistasis effect; aae_1 and aae_2: the environment-specific additive-additive epistasis effect; Pop, Power, Par., and Est. have same definitions as those in Table 1.

Based on these potential SNPs, the model (1) was applied for significance test of the genetic effects due to individual SNP and paired SNPs in terms of F-statistic and the threshold specified by the permutation method at the experiment-wise error rate of 0.05. Then, all the significant SNPs were used to build a full model and the MCMC method was employed to generate the distribution of each effect in the full model. On the basis of the distribution, each effect was estimated by the mean and its significance is tested by t-statistic [31]. The data analysis by the two-step strategy was implemented with a newly developed GWAS software called QTXNetwork (http://ibi.zju.edu.cn/software/QTXNetwork/).

3. Results

3.1. Monte Carlo Simulations. To investigate the efficiency and accuracy of the proposed methods, we performed a series of Monte Carlo simulations to verify the unbiasedness and robustness as well as statistical power. Because our real experiment data is based on a rice RIL population, we conducted the simulations based on this kind of population to examine our methods. Two environments were considered. 525 SNPs were scattered across 3 chromosomes, with 175 markers evenly distributed on each chromosome. Three individual QTSs controlling the quantitative trait were assigned on 3 chromosomes and two-paired epistatic QTSs were also set with gene-gene-environment interaction effects. Our simulations would also investigate the influences of heritability and population size on estimation of QTS parameters. Different heritabilities and population sizes were used in simulations. Each simulation generated the experimental samples according to the parameter setting for analysis and the results from 200 simulations are summarized in Tables 1–4.

Under a heritability of 70%, we conducted simulations to investigate the effectiveness and robustness of our proposed method in estimating QTS parameters, according to the results in Tables 1 and 2, it was clear that all the QTSs could be detected, and all kinds of effects could be estimated effectively under the three population sizes. Most estimates were close to true values of parameter, and the estimation accuracy of QTS effects was acceptable, especially for the QTS main effects. The estimates of additive-additive interaction effects with environments were less accurate, which might be due to the interference additive-additive effects. The statistical power increased as the population size became bigger and most of the power under the population size over 150 individuals was 100% except SNP.93 on the third chromosome.

In order to understand the robustness of our method, we conducted simulations under two heritabilities of 70% and 50%, respectively, under the population size of 150 individuals. The simulation results (Tables 3 and 4) clearly showed a very accurate detection of QTSs and estimation of QTS effects. All the estimates of QTS positions and effects were quite robust for different heritabilities, but the epistatic interaction effects with environments between SNP.44 on the second chromosome and SNP.63 on the third chromosome deviated a little from the true value. Additionally, the power did not changed very much when the heritability was increased from 50% to 70%.

Table 3: Estimates of SNP additive effects and interaction effects with the environments under two different heritabilities.

Chr	SNPID	a			ae_1			ae_2			Power	
		Par.	Est.		Par.	Est.		Par.	Est.			
			I	II		I	II		I	II	I	II
1	28	−0.79	−0.81	−0.83	0.63	0.65	0.67	−0.63	−0.60	−0.58	100.00	100.00
2	100	−0.67	−0.67	−0.68	0.39	0.38	0.39	−0.39	−0.37	−0.37	100.00	100.00
3	93	−0.40	−0.41	−0.41	0.32	0.32	0.32	−0.32	−0.31	−0.31	69.00	70.00

I and II stand for two different heritabilities of 50% and 70%, respectively, which are the proportions of total phenotypic variation ascribed to SNP additive effects and additive-environment interaction effects; Power, Par., Est., a, ae_1, and ae_2 have the same definitions as those in Table 1.

Table 4: Estimates of epistasis and interaction effects with the environments under two different heritabilities.

Chr.i	SNPID.i	Chr.j	SNPID.j	aa			aae_1			aae_2			Power	
				Par.	Est.		Par.	Est.		Par.	Est.			
					I	II		I	II		I	II	I	II
2	44	3	63	0.39	0.38	0.38	0.17	0.18	0.17	−0.17	−0.18	−0.16	100.00	99.50

Power, Par., and Est. have the same definitions as those in Table 1; aa, aae_1, and aae_2 have the same definitions as those in Table 2.

In conclusion, according to the increase of population size, the power values of QTS with median effects became higher and the estimated effects were closer to the true parameter values. Therefore, a population consisting of around 150 or more RILs is a reasonable size to maintain the estimation efficiency and power for a trait controlled by modest QTS effects.

3.2. The Genetic Architecture of Rice Quality Traits. Based on the above simulation consequences, an association study was conducted to analyze the genetic architecture for three quality traits of super-hybrid rice Xieyou9308 RIL population with 138 lines in two environments and the results were listed in Table 5. Totally, four QTSs were detected on all the rice chromosomes for the three quality traits, 2 QTSs for AC, 2 QTSs for GC, and 2 QTSs for GT, respectively. One QTS seemed to be pleiotropic for all the three traits. The total heritability estimated by the full model for the three traits was all over 50%, of which the highest was 68.67% for AC and the lowest was 52.01% for GT.

For AC, two QTSs were detected and involved only in additive effects with very high significance and both of effects were negative. Within them, the heritability of rs1644460 was as high as 51.78% whose proportion was up to 75.4% of total heritability.

There were also two QTSs discovered for GC. Positive additive effect of rs1644460 and negative additive effect of rs919289 were detected, similar to AC; the former, with an extremely high significance, accounted for more than 89% of the genetic variance of GC. The special QTS, rs1644460, was significant in both environments but showed varying effects in different environments, indicating an unstable expression of this QTS across different environments.

For GT, two significant QTSs were identified, which were involved not only in additive effects and additive × environment interaction effects but also in epistasis and interaction of epistasis with environments. The QTS, rs1289107, with highest additive effect reached the highest heritability (24%) of all the effects. Similar to GC, there were two QTSs, rs1289107 and rs1644460, which were expressed in distinct patterns under different environments and they reached high heritability in interaction effects of additive with environments, indicating that the expression of these two QTSs depended substantially on environments. It should be noted that one-paired epistatic QTSs were detected for GT with different pattern epistasis × environment interaction effects.

Throughout the three traits, rs1644460 was detected for all the three quality traits with high significance and heritability, suggesting a pleiotropic role of the QTS for the three traits. Generally, there were diverse patterns in genetic effects of QTSs among three quality traits. The QTSs controlling all the three traits expressed mainly in genetic main effects and subtle environment-specific additive or additive-additive epistasis effects were detected. Furthermore, only two highly significant QTSs (rs1644460 and rs1289107) were expressed both in genetic main effects and genetic-environment interaction effects. On the contrary, two QTSs (rs1610021 and rs919289) were expressed mainly in additive effects and modestly in genetic-environment interaction effects for AC and GC. Conclusively, environment is a very crucial factor to affect gene expression for rice quality traits. Most QTSs interact with environment and the environment can enhance or weaken the expression of most QTSs for the three quality traits.

4. Discussion

Epistasis and its interaction with environment have been recognized as important components of cultivar performance and have received more attention in rice breeding programs. Nevertheless, these effects have not yet effectively been analyzed by the current GWASs for the reason of absence of appropriate statistical methodology and heavy burden of computation [32–34]. If a reduced model ignoring epistasis and gene-environment interaction is employed, the resulting GWAS would give biased estimation of effects, poor detection

Table 5: Detected SNPs with significant genetic effects.

Trait	QTS	Chr.	Allele	Effect	Predict	$-\log_{10}(P)$	h^2 (%)	Total
AC	rs1610021	6	G/A	a	−2.20	26.94	16.89	68.67
	rs1644460	6	C/T	a	−3.85	79.97	51.78	
GC	rs1644460	6	C/T	a	15.02	61.32	52.15	58.58
				ae_1	−3.48	2.52	2.80	
				ae_2	3.47	2.11	2.80	
	rs919289	7	C/G	a	−3.96	4.97	3.63	
GT	rs1289107	6	G/A	a	−0.54	25.06	24.00	52.01
				ae_1	0.25	3.49	5.00	
				ae_2	−0.25	2.99	5.00	
	rs1644460	6	C/T	a	−0.35	10.73	9.81	
				ae_1	−0.30	5.01	7.40	
				ae_2	0.30	4.05	7.40	
	rs1289107 and rs1644460	6	G/A	aa	0.21	4.33	3.60	
				aae_1	−0.16	1.76	2.20	
		6	C/T	aae_2	0.17	1.61	2.20	

AC: amylose content; GC: gel consistency; GT: gelatinization temperature; e_1: environment 1; e_2: environment 2; a: additive effect; ae_1 and ae_2: environment-specific additive effect; aae_1 and aae_2: environment-specific additive-additive epistasis effect. $-\log_{10}(P) = -\log_{10}(P\text{-value})$. h^2 (%) = heritability (%) due to the genetic component effect.

precision and power, and low heritability to explain variation of complex traits [34–37]. This study used a QTS full model including the additive effect, the additive-additive epistatic effect, and their interaction effects with multienvironments of each QTS, to analyze genetic architecture of gelatinization temperature, amylose content, and gel consistency so that the estimation accuracy of each effect will be greatly improved benefiting from elimination of false positive QTS by permutation method and more accurate estimation of residual effect. In order to alleviate the computing cost, we first used the GMDR algorithm on GPU to screen the potential associated markers and then conducted one-dimensional and two-dimensional searching to detect putative QTSs with the screened markers as cofactors to control background genetic effects. Totally, we identified four QTSs with additive effects, epistasis, and environment interaction effects for the three quality traits.

Traditional plant breeding is based on phenotypic selection of superior genotypes among segregating progenies, and its effectiveness is often affected by environment and genotype-environment interaction. Therefore, it sometimes leads to unreliable selection of some traits [38–40]. Although marker-assisted selection (MAS) is an effective way to improve the efficiency and precision of plant breeding, it is still under the influence of the strength in association between markers and genes of target traits. As the highly significant SNPs identified by GWAS mapping mostly link tightly to the genes controlling target traits, with assistance of these detected QTSs in this study, selection of the quality traits will be more efficient and accurate improvement of target traits will be fast achieved. Further, based on the information of genetic main effects of QTSs or interaction effects of QTSs with environments, it becomes possible for us to design an universal selection strategy effective for all the environments or a specific selection strategy for individual environments.

According to the association analysis, two QTSs (rs1610021 and rs1289107) of the four QTSs, which are involved in the genetic variation of three quality traits of rice, are located in the region of the known genes (Os06g0130100 and Os06g0124300), and the other two, rs1644460 and rs919289, lie in the upstream or the downstream of known genes (Os06g0130800 and Os07g0116900). Some of these genes have been well defined; for example, there is one QTS, rs919289, near the gene of Os07g0116900 that is described as NADH ubiquinone oxidoreductase, 20 kDa subunit domain containing protein. NADH plays essential roles in metabolism, which emerges as an adenine nucleotide that can be released from cells spontaneously and by regulated mechanisms [41, 42]. But the function of the others or the relationship between the gene and the three quality traits of rice still remains unexplored, such as Os06g0130100 whose annotation is similar to ERECTA-like kinase 1. It has been previously reported that ERECTA-family receptor-like kinases (RLKs) are redundant receptors that relate cell proliferation to organ growth and patterning [43]. Further investigation is needed to explain the association between RLK ERECTA and the three quality traits of rice.

Although the three rice quality traits exhibit different genetic architecture in the pattern of genetic effects, we believe that all of these are crucial genetic resources for improvement of the traits by selection of genetic effects. However, more validation is needed if these QTSs will be extended to other populations with different genetic background. In addition, more detailed whole-genome scanning, more powerful bioinformatics tools, and larger size of mapping

populations are required since only causative polymorphisms with large effects can be detected given the size of the used RIL population [44].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This study was supported in part by the Project of the National Sci-Tech Support Plan (2011BAD35B02), the Major Sci-Tech Programs of Zhejiang Province (2012C12901-3), the National Natural Science Foundation Grants of China 31271608 and 31470083, the National Institutes of Health Grant DA025095, and the Bill and Melinda Gates Foundation Project.

References

[1] S.-H. Cheng, L.-Y. Cao, J.-Y. Zhuang et al., "Super hybrid rice breeding in China: achievements and prospects," *Journal of Integrative Plant Biology*, vol. 49, no. 6, pp. 805–810, 2007.

[2] M. Benmoussa, K. A. K. Moldenhauer, and B. R. Hamaker, "Rice amylopectin fine structure variability affects starch digestion properties," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 4, pp. 1475–1479, 2007.

[3] J. W. Cone and M. G. Wolters, "Some properties and degradability of isolated starch granules," *Starch—Stärke*, vol. 42, no. 8, pp. 298–301, 1990.

[4] S. J. Evans, P. V. Choudary, C. R. Neal et al., "Dysregulation of the fibroblast growth factor system in major depression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 43, pp. 15506–15511, 2004.

[5] J. H. Li, T. Vasanthan, R. Hoover, and B. G. Rossnagel, "Starch from hull-less barley: V. In-vitro susceptibility of waxy, normal, and high-amylose starches towards hydrolysis by alpha-amylases and amyloglucosidase," *Food Chemistry*, vol. 84, no. 4, pp. 621–632, 2004.

[6] C. A. Rendleman, V. E. Beckner, M. Lijewski, W. Crutchfield, and J. B. Bell, "Parallelization of structured, hierarchical adaptive mesh refinement algorithms," *Computing and Visualization in Science*, vol. 3, no. 3, pp. 147–157, 2000.

[7] C. K. Riley, A. O. Wheatley, I. Hassan, M. H. Ahmad, E. Y. S. A. Morrison, and H. N. Asemota, "*In vitro* digestibility of raw starches extracted from five yam (*Dioscorea* spp.) species grown in jamaica," *Starch/Stärke*, vol. 56, no. 2, pp. 69–73, 2004.

[8] Y. Sang, S. Bean, P. A. Seib, J. Pedersen, and Y.-C. Shi, "Structure and functional properties of sorghum starches differing in amylose content," *Journal of Agricultural and Food Chemistry*, vol. 56, no. 15, pp. 6680–6685, 2008.

[9] H. Themeier, J. Hollmann, U. Neese, and M. G. Lindhauer, "Structural and morphological factors influencing the quantification of resistant starch II in starches of different botanical origin," *Carbohydrate Polymers*, vol. 61, no. 1, pp. 72–79, 2005.

[10] T. Vasanthan and R. S. Bhatty, "Physicochemical properties of small- and large-granule starches of waxy, regular, and high-amylose barleys," *Cereal Chemistry*, vol. 73, no. 2, pp. 199–207, 1996.

[11] N. Asp, *Resistant Starch: Proceedings from the 2nd Plenary Meeting of EURESTA: European Flair Concerted Action no. 11 (COST 911): Physiological Implications of the Consumption of Resistant Starch in Man*, Macmillan, 1992.

[12] G. S. Khush, C. M. Paule, and N. M. De La Cruz, "Rice grain quality evaluation and improvement at IRRI," in *Proceedings of the Workshop on Chemical Aspects of Rice Grain Quality*, pp. 21–31, Los Baños, Philippines, 1978.

[13] B. O. Juliano, "The chemical basis of rice grain quality," in *Proceedings of the Workshop on Chemical Aspects of Rice Grain Quality*, pp. 69–90, International Rice Research Institute, Los Baños, Philippines, 1979.

[14] X. Huang, X. Wei, T. Sang et al., "Genome-wide asociation studies of 14 agronomic traits in rice landraces," *Nature Genetics*, vol. 42, no. 11, pp. 961–967, 2010.

[15] K. L. Kump, P. J. Bradbury, R. J. Wisser et al., "Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population," *Nature Genetics*, vol. 43, no. 2, pp. 163–168, 2011.

[16] J. A. Poland, P. J. Bradbury, E. S. Buckler, and R. J. Nelson, "Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 17, pp. 6893–6898, 2011.

[17] J. A. Rafalski, "Association genetics in crop improvement," *Current Opinion in Plant Biology*, vol. 13, no. 2, pp. 174–180, 2010.

[18] J. M. Thornsberry, M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. S. Buckler, "Dwarf8 polymorphisms associate with variation in flowering time," *Nature Genetics*, vol. 28, no. 3, pp. 286–289, 2001.

[19] F. Tian, P. J. Bradbury, P. J. Brown et al., "Genome-wide association study of leaf architecture in the maize nested association mapping population," *Nature Genetics*, vol. 43, no. 2, pp. 159–162, 2011.

[20] W. A. Cowling and E. Balázs, "Prospects and challenges for genome-wide association and genomic selection in oilseed Brassica species," *Genome*, vol. 53, no. 11, pp. 1024–1028, 2010.

[21] R. Dobson, S. V. Ramagopalan, and G. Giovannoni, "Genome-wide association studies: will we ever predict susceptibility to multiple sclerosis through genetics?" *Expert Review of Neurotherapeutics*, vol. 13, no. 3, pp. 235–237, 2013.

[22] C. E. Murcray, J. P. Lewinger, and W. J. Gauderman, "Gene-environment interaction in genome-wide association studies," *American Journal of Epidemiology*, vol. 169, no. 2, pp. 219–226, 2009.

[23] Z. Zhu, X. Tong, M. Liang et al., "Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes," *PLoS ONE*, vol. 8, no. 4, Article ID e61943, 2013.

[24] Z.-K. Li, L. J. Luo, H. W. Mei et al., "Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield," *Genetics*, vol. 158, no. 4, pp. 1737–1753, 2001.

[25] H. W. Mei, Z. K. Li, Q. Y. Shu et al., "Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two backcross populations," *Theoretical and Applied Genetics*, vol. 110, no. 4, pp. 649–659, 2005.

[26] A. Q. You, X. Lu, H. Jin et al., "Identification of quantitative trait loci across recombinant inbred lines and testcross populations for traits of agronomic importance in rice," *Genetics*, vol. 172, no. 2, pp. 1287–1300, 2006.

[27] Y. Kawahara, M. de la Bastide, J. P. Hamilton et al., "Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, article 4, 2013.

[28] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[29] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[30] G.-B. Chen, Y. Xu, H.-M. Xu, M. D. Li, J. Zhu, and X.-Y. Lou, "Practical and theoretical considerations in study design for detecting gene-gene interactions using MDR and GMDR approaches," *PLoS ONE*, vol. 6, no. 2, Article ID e16981, 2011.

[31] J. Yang, J. Zhu, and R. W. Williams, "Mapping the genetic architecture of complex traits in experimental populations," *Bioinformatics*, vol. 23, no. 12, pp. 1527–1536, 2007.

[32] Ö. Carlborg and C. S. Haley, "Epistasis: too often neglected in complex trait studies?" *Nature Reviews Genetics*, vol. 5, no. 8, pp. 618–625, 2004.

[33] P. C. Phillips, "Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems," *Nature Reviews Genetics*, vol. 9, no. 11, pp. 855–867, 2008.

[34] J. van Os and B. P. F. Rutten, "Gene-environment-wide interaction studies in psychiatry," *The American Journal of Psychiatry*, vol. 166, no. 9, pp. 964–966, 2009.

[35] B. Brachi, G. P. Morris, and J. O. Borevitz, "Genome-wide association studies in plants: the missing heritability is in the field," *Genome Biology*, vol. 12, no. 10, article 232, 2011.

[36] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich, "A perspective on epistasis: limits of models displaying no main effect," *The American Journal of Human Genetics*, vol. 70, no. 2, pp. 461–471, 2002.

[37] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 4, pp. 1193–1198, 2012.

[38] E. Francia, G. Tacconi, C. Crosatti et al., "Marker assisted selection in crop plants," *Plant Cell, Tissue and Organ Culture*, vol. 82, no. 3, pp. 317–342, 2005.

[39] B. Holloway and B. Li, "Expression QTLs: applications for crop improvement," *Molecular Breeding*, vol. 26, no. 3, pp. 381–391, 2010.

[40] M. Mohan, S. Nair, A. Bhagwat et al., "Genome mapping, molecular markers and marker-assisted selection in crop plants," *Molecular Breeding*, vol. 3, no. 2, pp. 87–103, 1997.

[41] R. A. Billington, S. Bruzzone, A. de Flora et al., "Emerging functions of extracellular pyridine nucleotides," *Molecular Medicine*, vol. 12, no. 11-12, pp. 324–327, 2006.

[42] L. M. Smyth, J. Bobalova, M. G. Mendoza, C. Lew, and V. N. Mutafova-Yambolieva, "Release of β-nicotinamide adenine dinucleotide upon stimulation of postganglionic nerve terminals in blood vessels and urinary bladder," *The Journal of Biological Chemistry*, vol. 279, no. 47, pp. 48893–48903, 2004.

[43] E. D. Shpak, C. T. Berthiaume, E. J. Hill, and K. U. Torii, "Synergistic interaction of three ERECTA-family receptor-like kinases controls *Arabidopsis* organ growth and flower development by promoting cell proliferation," *Development*, vol. 131, no. 7, pp. 1491–1501, 2004.

[44] B. Peng, Y. Li, Y. Wang et al., "QTL analysis for yield components and kernel-related traits in maize across multi-environments," *Theoretical and Applied Genetics*, vol. 122, no. 7, pp. 1305–1320, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 916352, 13 pages
http://dx.doi.org/10.1155/2015/916352

Research Article

Dynamic Model for RNA-seq Data Analysis

Lerong Li and Momiao Xiong

Human Genetics Center, Division of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Correspondence should be addressed to Momiao Xiong; momiao.xiong@uth.tmc.edu

Received 4 December 2014; Accepted 16 February 2015

Academic Editor: Ernesto Picardi

By measuring messenger RNA levels for all genes in a sample, RNA-seq provides an attractive option to characterize the global changes in transcription. RNA-seq is becoming the widely used platform for gene expression profiling. However, real transcription signals in the RNA-seq data are confounded with measurement and sequencing errors and other random biological/technical variation. To extract biologically useful transcription process from the RNA-seq data, we propose to use the second ODE for modeling the RNA-seq data. We use differential principal analysis to develop statistical methods for estimation of location-varying coefficients of the ODE. We validate the accuracy of the ODE model to fit the RNA-seq data by prediction analysis and 5-fold cross validation. To further evaluate the performance of the ODE model for RNA-seq data analysis, we used the location-varying coefficients of the second ODE as features to classify the normal and tumor cells. We demonstrate that even using the ODE model for single gene we can achieve high classification accuracy. We also conduct response analysis to investigate how the transcription process responds to the perturbation of the external signals and identify dozens of genes that are related to cancer.

1. Introduction

Next-generation sequencing (NGS) technologies have revolutionized advances in the study of the transcriptome. The newly developed deep-sequencing technologies make it possible to acquire both quantitative and qualitative information regarding transcript biology. By measuring messenger RNA levels for all genes in a sample, RNA-seq provides an attractive option to characterize the global changes in transcription.

To generate RNA-seq data, the complete set of mRNAs are first extracted from an RNA sample and then shattered and reverse-transcribed into a library of cDNA fragments with adaptors attached. These short pieces of cDNA are amplified by polymerase chain reaction and sequenced by machine, producing millions of short reads. These reads are then mapped to a reference genome or reference transcript. The number of reads within a region of interest is used as a measure of abundance. The reads can also be assembled de novo without the genomic sequence to create a transcription map.

Compared to microarray which provides limited gene regulation information, RNA-seq offers a comprehensive picture of the transcriptome. RNA-seq has made a number of significant qualitative and quantitative improvements on gene expression analysis and provides multiple layers of resolutions and transcriptome complexity: the expression at exon, SNP, and positional level; splicing; posttranscriptional RNA editing across the entire gene; isoform and allele-specific expression [1].

Many advantages include strong concordance between platforms, higher sensitivity and dynamic range, lower technical variation and background signal, and high level of technical and biological reproducibility, and so on [2–5]. However, some limitations are inherent to next-generation sequencing technology. For example, the read coverage may not be homogeneous along the genome, and different samples may be sequenced at different levels of depth in the experiment. Also, although some genes may have a similar level of expression, longer genes are more likely to have more reads than short ones. Therefore, RNA-seq data must be normalized before any comparison of the counts can be made. Another consideration is that, in production of cDNA libraries, larger RNA must be fragmented into smaller pieces to be sequenced and different fragmentations may create bias towards different outcomes. Some other informatics challenges like the storage, transfer, and retrieval of large size data may bring additional errors [6, 7].

Expression variability measured by RNA-seq arises from three primary sources: (i) real biological differences in different experimental groups or conditions, (ii) measurement errors, and (iii) random biological and/or technical variation [1, 8]. The first type of variability is of real biological interest but is confounded with measurement and sequencing errors and other random biological/technical variation. How to appropriately take the latter two types of variability into account is essential issue in the RNA-seq data analysis.

The purpose of this paper is to borrow dynamic theory from engineering and use ordinary differential equation (ODE) for modelling the observed number of reads across the gene and unravelling the features of gene transcription [9]. To achieve this goal, we considered the number of reads or expression level at each position as a function of the genomic position and viewed the transcription process as a stochastic process of transcription along the genome. Instead of taking the derivative of expression level with respect to time, we calculated the expression level derivative with respect to genomic position. Specifically, we proposed a dynamic model for the variation of the transcription process along the genome. For each gene, we use a second order ODE with location-dependent coefficient to model that gene's transcription process. We develop statistical methods for estimation of the coefficient functions in the ODE based on principal differential analysis. Compared to the ODE model with constant coefficient to capture the stochastic variation feature of transcription process, the location-dependent coefficients are essential to account for the complicated stochastic process of gene regulation.

To examine the precision of the ODE for modeling the RNA-seq data, we split the samples into five groups and use 5-fold cross validation to evaluate the accuracy of predicting gene expression level across the gene using the ODE model.

To capture stochastic feature of gene regulation, we conduct the response analysis. The response analysis of transcriptional processes for each gene using its fitted differential equation can provide important aspects of transcription, including alternative splicing, alternative start and end of transcription, and alternative isoforms. To differentiate feature of gene regulation between normal and cancer tissue samples, we develop statistics to test for significant difference in the response of the gene regulation between the normal and cancer samples under the perturbation of external signals and perform genome-wide response analysis of gene regulation. Using the ODE model, we identified the genes that have a significantly different transcriptional process (both different magnitude and different patterns) and identified genes that showed significantly differential stochastic behaviors in response to environmental perturbations between normal and cancer samples.

To further explore application of the ODE for RNA-seq data analysis, we take the location-varying coefficients of the ODE as features and use FPCA as a tool for extraction of these features. The FPCA scores are used as features and the Lasso logistic regression is used as feature selection tool and classifier for distinguishing the cancer and normal samples.

The data suggest that the dynamic features of gene transcription captured by the coefficient functions can retrieve the original process information. Therefore, they naturally served as good candidate's features for clustering genes with similar transcription process. These groups of genes could share common biological function, chromosomal location, pathway, or regulation. The ODE for modeling the RNA-seq data has the potential to provide valuable information for understanding the mechanism of gene regulation and unraveling disease processes.

2. Materials and Methods

2.1. ODE Model with Varying Coefficients for RNA-seq Data. Assume that the expression of a gene is measured by the number of sequence reads mapped to this gene in the region $T = [a, b]$. Let t denote a genomic position, let $y(t)$ be observed gene expression level that was measured by the number of reads mapped to the genomic position, and let $x(t)$ be the hidden state that determined the gene expression level at the genomic position t. To model transcription process, the second order ordinary differential equation (ODE) with location-varying coefficients can be specified as follows:

$$L(x(t)) = \frac{d^2x(t)}{dt^2} + w_1(t)\frac{dx(t)}{dt} + w_0(t)x(t) = 0, \quad (1)$$

where $w_1(t)$ and $w_0(t)$ are weighting coefficients or parameters in the ODE. Its observations $y(t)$ often have measurement errors:

$$y(t) = x(t) + e(t), \quad (2)$$

where $e(t)$ is measurement error at the position t.

2.2. Estimation of Coefficient Functions in the ODE. Estimation of coefficient functions in the ODE consists of two steps. At the first step we estimate the states $\hat{x}(t)$ from the observed number of reads $y(t)$ assuming that coefficient functions in the ODE are given. At the second step, we estimate the coefficient functions in the ODE, assuming that states $x(t)$ have been estimated.

Step 1. To estimate $x(t)$, we first expand the function $x(t)$ in terms of basis functions $\phi(t)$ and then estimate its expansion coefficients. Let $x_i(t)$ be the state variable at the genomic position t of the ith sample and let $y_i(t)$ be its observation ($i = 1, \ldots, n$). Then, $x_i(t)$ can be expanded as

$$x_i(t) = \sum_{j=1}^{K} c_{ij}\phi_j(t) = C_i^T\phi(t), \quad (3)$$

where $C_i = [c_{i1}, \ldots, c_{iK}]^T$ and $\phi(t) = [\phi_1(t), \ldots, \phi_K(t)]^T$.

Similarly, the parameters $w_1(t)$ and $w_0(t)$ can be expanded as

$$w_1(t) = \sum_{j=1}^{K} h_{1j}\phi_j(t) = h_1^T\phi(t),$$
$$w_0(t) = \sum_{j=1}^{K} h_{0j}\phi_j(t) = h_0^T\phi(t). \quad (4)$$

Substituting their expansions into (1), we obtain

$$L\left(x_i\left(t\right)\right) = C_i^T \Psi\left(t\right), \tag{5}$$

where $\Psi(t) = d^2\phi/dt^2 + G(t)h$, $G(t) = [(d\phi/dt)\phi^T(t), \phi(t)\phi^T(t)]$, and $h = [h_1^T, h_0^T]^T$. To smooth the estimated function $\widehat{x}(t)$, we impose the following penalty term:

$$\boldsymbol{\lambda} \int_T L\left(x_i\left(t\right)\right) L^T\left(x_i\left(t\right)\right) dt = \boldsymbol{\lambda} C_i^T J_{\phi h} C_i, \tag{6}$$

where $J_{\phi h} = \int_T \Psi(t)\Psi^T(t)dt$.

We estimate the state function $x(t)$ from the observation data $y(t)$ by minimizing the following objective function which consists of the sum of the squared errors between the observations and the estimated states and the penalty terms:

$$\begin{aligned}&\sum_{i=1}^{n}\left\{\sum_{j=1}^{\tau}\left[y_i\left(t_j\right) - x_i\left(t_j\right)\right]^2 + \boldsymbol{\lambda}\int_T L\left(x_i\left(t\right)\right) L^T\left(x_i\left(t\right)\right) dt\right\}\\ &\quad = \sum_{i=1}^{n}\left\{\sum_{j=1}^{\tau}\left[y_i\left(t_j\right) - x_i\left(t_j\right)\right] + \boldsymbol{\lambda} C_i^T J_{\phi h} C_i\right\}.\end{aligned} \tag{7}$$

Let

$$\begin{aligned}&Y_i = \left[y_i\left(t_1\right), \ldots, y_i\left(t_T\right)\right]^T, \qquad Y = \left[Y_1^T, \ldots, Y_n^T\right],\\ &\widetilde{\phi} = \left[\phi\left(t_1\right), \ldots, \phi\left(t_T\right)\right]^T, \qquad C = \left[C_1^T, \ldots, C_n^T\right]^T,\\ &\Phi = \operatorname{daig}\left(\widetilde{\phi}, \ldots, \widetilde{\phi}\right), \qquad J = \operatorname{diag}\left(J_{\varphi h}, \ldots, J_{\phi h}\right).\end{aligned} \tag{8}$$

Problem (7) can be rewritten in a matrix form:

$$\min_C \left(Y - \Phi C\right)^T \left(Y - \Phi C\right) + \boldsymbol{\lambda} C^T J C. \tag{9}$$

The least square estimators of the expansion coefficients are then given by

$$C = \left(\Phi^T \Phi + \boldsymbol{\lambda} J\right)^{-1} \Phi. \tag{10}$$

Step 2. Next we estimate the coefficient functions in the ODE. The coefficient functions in the ODE can be estimated by minimizing the following least squares objective function:

$$\min_h \mathrm{SSE}_p = \int_T L^T\left(\widehat{X}\left(t\right)\right) L\left(\widehat{X}\left(t\right)\right) dt, \tag{11}$$

where $L(\widehat{X}(t)) = [L(\widehat{x}_1(t)), \ldots, L(\widehat{x}_n(t))]^T$.

Since $L(x_i(t)) = C_i^T \Psi(t)$, the $L(x(t))$ can be expressed in terms of the estimated expansion coefficients as

$$L\left(\widehat{X}\left(t\right)\right) = C_* \Psi\left(t\right), \tag{12}$$

where the matrix C_* is defined as

$$C_* = \begin{bmatrix} C_1^T \\ \vdots \\ C_n^T \end{bmatrix}. \tag{13}$$

Therefore, problem (11) can be reduced as

$$\min_h \mathrm{SSE}_p = \int_T \Psi^T\left(t\right) C_*^T C_* \Psi\left(t\right) dt, \tag{14}$$

where the matrix C_* is estimated and hence fixed in minimization problem (14). Setting the partial derivative of SSE_p to be zero,

$$\frac{\partial \mathrm{SSE}_p}{\partial h} = \int_T G^T\left(t\right) C_*^T C_* \left[\frac{d^2\phi}{dt^2} + G\left(t\right) h\right] dt = 0. \tag{15}$$

Solving (15) for h, we obtain

$$h = -\left[\int_T G^T\left(t\right) C_*^T C_* G\left(t\right) dt\right]^{-1} \int_T G^T\left(t\right) C_*^T C_* \frac{d^2\phi}{dt^2} dt. \tag{16}$$

In summary, we iteratively determine the expansion coefficients of the state function $X(t)$ for fixed parameters in the ODE by (10) and estimate the coefficient functions in the ODE for fixed expansion coefficients by (16).

2.3. ODE for Classification. To illustrate that the ODE can be used as a useful tool for modeling the profile of the RNA-seq expression we will show that the ODE can capture all variation of gene expression across the gene and that the coefficient functions of the ODE are useful feature extraction of the RNA-seq data. The ODE can be used for classifying tumor and normal samples.

Since dimensions of the coefficient functions of the ODE are extremely high, the functional principal component analysis (FPCA) is used to reduce the dimensions of the coefficient functions of the ODE.

The FPCA tries to find the dominant direction of variation around an overall trend function [10, 11]. Each principal component is specified by the weight function $\beta(t)$, and the principal component scores of the individuals in the sample are defined as the inner product of weight function and functional curves $(w_0(t), w_1(t))$:

$$z = \int_T \beta\left(t\right) w\left(t\right) dt, \tag{17}$$

where, for convenience, we use $w(t)$ to denote either $w_1(t)$ or $w_0(t)$, that is, the coordinate value of functional curves at the direction of $\beta(t)$ with highest variability. By projecting the functional curves onto set of eigenfunctions, we can reduce the dimension to finite number, functional principal component scores.

Suppose that for the ith individual sample we obtain the functional principal component score:

$$\begin{aligned}z_{ij}^{(1)} &= \int_T \beta_j\left(t\right) w_{i1}\left(t\right) dt,\\ z_{ij}^{(0)} &= \int_T \beta_j\left(t\right) w_{i0}\left(t\right) dt,\end{aligned} \tag{18}$$

where $w_{i1}(t)$ and $w_{i0}(t)$ are the coefficient functions of the ODE for the ith individual sample and $\beta_j(t)$, $j = 1, \ldots, K$, are

a set of eigenfunctions (or principal component functions). The original functional curves can be reduced to a finite feature matrix:

$$Z = \begin{bmatrix} z_{11}^{(0)} & z_{11}^{(1)} & \cdots & z_{1K}^{(0)} & z_{1K}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ z_{n1}^{(0)} & z_{n1}^{(1)} & \cdots & z_{nK}^{(0)} & z_{nK}^{(1)} \end{bmatrix}, \tag{19}$$

where the K is the number of principal components selected to explain the total variability.

To improve classification accuracy we use the Lasso logistic regression as a classifier. In simple logistic regression, we use the logit link to relate the mean of response with the covariates of interest. Let $\mathbf{x}_i = [x_1, \ldots, x_p]^T$ be the vector of observed covariates for ith observation, and y_i is the corresponding response outcome. For simplicity, we consider binary cases where $y_i = 1$ or 0. The model is specified as the following posterior probability for ith observation [12]:

$$\pi_i(\mathbf{x}_i, \boldsymbol{\beta}) = \Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp\left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right)}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right)}, \tag{20}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is the covariate vector of interest and β_0 is the intercept term. And the joint log-likelihood of the N subjects is defined as

$$l(\beta) = \sum_{i=1}^{N} \log \pi_i(\mathbf{x}_i, \boldsymbol{\beta}) \tag{21}$$

which can be written as

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \{y_i \log \pi_i(\mathbf{x}_i, \boldsymbol{\beta}) + (1 - y_i) \log(1 - \pi_i(\mathbf{x}_i, \boldsymbol{\beta}))\} \\ &= \sum_{i=1}^{N} \left\{ y_i \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right) - \log\left(1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}\right) \right\}. \end{aligned} \tag{22}$$

To estimate the parameter, we set its derivatives to zero and get the score equations

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} \mathbf{x}_i (y_i - \pi_i(\mathbf{x}_i, \beta)) = 0. \tag{23}$$

Since (23) is nonlinear equation in β, we usually use some iterative methods like Newton-Raphson algorithm to get the solution of β.

By adding an L_1 penalty to the joint log-likelihood in (23) we have the following constrained maximization equation:

$$\begin{aligned} l_c(\boldsymbol{\beta}) = \Bigg\{ &\sum_{i=1}^{N} \left\{ y_i \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right) - \log\left(1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}\right) \right\} \\ &- \boldsymbol{\lambda} \sum_{j=1}^{p} |\beta_j| \Bigg\}, \end{aligned} \tag{24}$$

where $l_c(\boldsymbol{\beta})$ is the constrained log-likelihood and $\boldsymbol{\lambda}$ is tuning parameter to adjust the tradeoff between log-likelihood function and the size of penalty. Please note that, in Lasso, we usually do not penalize the intercept term and it is practically meaningful to standardize the covariates before optimization.

The L_1 penalty is not differentiable and also $\boldsymbol{\beta}$ is not linear solution of response $\mathbf{y}$. It is not trivial to get the score functions but we can still have a solution using nonlinear programming method [13]. The score functions for variables with nonzero coefficients have the form

$$\mathbf{x}_j^T (\mathbf{y} - \boldsymbol{\pi}) = \boldsymbol{\lambda} \cdot \operatorname{sign}\left(\beta_j\right). \tag{25}$$

Coordinate descent method is one efficient method to compute the Lasso solution. It fixes the penalty parameters $\boldsymbol{\lambda}$ and optimize over each parameter successively, while holding the others fixed at current values. R package *glmnet* [14] can efficiently fit the Lasso logistic regression with large N and p.

2.4. Numerical Solution to the ODE with Bounded Values. We use collocation Runge-Kutta method for the solution of boundary value problem of the ODE. The basic idea is to find a set of polynomials $p_n(x)$ of degree s which satisfies the problem over the interval $[x_{n-1}, x_n]$ for a set of points

$$x_{nj} = x_{n-1} + a_j h_n, \quad \text{where } j = 1, \ldots, s, \ n = 1, \ldots, N. \tag{26}$$

Note that, $0 < a_0 < a_1 < \cdots < a_s < 1$, they are distinct real numbers. Also the polynomial functions $p_n(x)$ are set to satisfy

$$\begin{aligned} &p_n(x_{n-1}) = y_{n-1}, \\ &p_n'(x_{nj}) = f\left(x_{nj}, p\left(x_{nj}\right)\right), \\ &\qquad \text{where } j = 1, \ldots, s. \end{aligned} \tag{27}$$

The numerical approximation at x_n is given by

$$y_n = p_n(x_{n-1} + h_n). \tag{28}$$

R package *bvpSolve* implements the method for boundary value problem [15].

2.5. Response Analysis under Perturbation of External Signal. Gene regulatory properties are encoded in the parameter curves of the ODE modeling gene expressions. Testing significant difference in the parameter curves between two conditions can be used as a powerful tool to assess differential changing behaviors of the gene expression across the gene region between two conditions. Response analysis attempts to extract inherent features of the systems that capture and describe the behaviors of the system over genomic positions under different operating conditions and perturbation of external signals.

Let t denote a genomic region within the gene of interests and let $x(t)$ be the number of reads mapped to the genomic region. And the ODE model used to describe the expression profile is given as follows:

$$L(x(t)) = \frac{d^2 x(t)}{dt^2} + w_1(t) \frac{dx(t)}{dt} + w_0(t) x(t) = 0. \tag{29}$$

Suppose the $\widehat{w_1}(t)$ and $\widehat{w_0}(t)$ are estimated from the data. The response of a regulatory system depends on the input signals. Different signal will cause different responses. For simplicity, we consider unit-step signal forced on the system and then solve the responses of the original system between different groups using estimated parameters $\widehat{w_1}(t)$ and $\widehat{w_0}(t)$:

$$\frac{d^2x(t)}{dt^2} + \widehat{w_1}(t)\frac{dx(t)}{dt} + \widehat{w_0}(t)x(t) = U(t). \tag{30}$$

To solve the solution of the estimated ODE with unit-step force function $U(t)$, we have to use some numerical methods to approximate the solution $\hat{x}(t)$. We solved ODE numerically by considering two-point boundary value problems where boundary conditions are specified at both ends of the range of integration. We estimated two initial values at both ends by evaluating the estimated smoothing expression curves at start and end positions.

Suppose $R(t) = [r_1(t), r_2(t), \ldots, r_{N_1}(t)]^T$ is a vector-valued function to represent response functional for all N_1 subjects in the normal group and $S(t) = [s_1(t), s_2(t), \ldots, s_{N_2}(t)]^T$ is response functional for N_2 subjects in cancer group. Therefore, we can construct a Hotelling T^2. Suppose that the response functions were expanded in terms of eigenfunctions $\phi_1(t), \ldots, \phi_K(t)$:

$$\begin{aligned} r_i(t) &= \sum_{j=1}^{K} \xi_{ij}\phi_j(t), \\ s_i(t) &= \sum_{j=1}^{K} \eta_{ij}\phi_j(t), \end{aligned} \tag{31}$$

where $\xi_{ij} = \int_T r_i(t)\phi_j(t)dt$ and $\eta_{ij} = \int_T s_i(t)\phi_j(t)dt$ and ξ_{ij} and η_{ij} are uncorrelated random variables with zero mean and variances $\boldsymbol{\lambda}_j$ with $\sum_j \boldsymbol{\lambda}_j < \infty$. Define the averages $\overline{\xi}_j$ and $\overline{\eta}_j$ of the principal component scores ξ_{ij} and η_{ij} in the normal and cancer group. Then we denote the average vector of scores in normal and cancer group by

$$\begin{aligned} \overline{\xi} &= \left[\overline{\xi_1}, \ldots \overline{\xi_k}\right]^T, \\ \overline{\eta} &= \left[\overline{\eta_1}, \ldots \overline{\eta_k}\right]^T, \end{aligned} \tag{32}$$

where $\overline{\xi_J} = (1/N_1)\sum_{i=1}^{N_1}\xi_{ij}$ and $\overline{\eta_J} = (1/N_2)\sum_{i=1}^{N_2}\eta_{ij}$, $j = 1, \ldots, k$.

The pooled covariance matrix is

$$S = \frac{1}{N_1 + N_2 - 2} \cdot \left(\sum_{i=1}^{N_1}\left(\xi_i - \overline{\xi}\right)\left(\xi_i - \overline{\xi}\right)^T + \sum_{i=1}^{N_2}\left(\eta_i - \overline{\eta}\right)\left(\eta_i - \overline{\eta}\right)^T\right), \tag{33}$$

where $\xi_i = [\xi_{i1}, \ldots, \xi_{ik}]^T$ and $\eta_i = [\eta_{i1}, \ldots, \eta_{ik}]^T$.

Let $\Lambda = (1/N_1 + 1/N_2)S$; then the Hotelling statistics can be written as

$$T^2 = \left(\overline{\xi} - \overline{\eta}\right)\Lambda^{-1} = \left(\overline{\xi} - \overline{\eta}\right)^T. \tag{34}$$

Under null of no difference in the response of the gene regulation between two groups, the statistics follows χ_k^2 distribution where k is the number of principle component scores.

3. Results

3.1. Dataset. We apply the proposed model to kidney renal clear cell carcinoma (KIRC) RNA-seq data, which is available from The Cancer Genome Atlas (TCGA) project (https://tcga-data.nci.nih.gov/tcga/). The RNA-seq data is available for 72 matched pairs of KIRC and normal samples. The maximum number of genomic positions where the expressions were measured by the number of reads passing quality control is 382,239,893 in the raw BAM file. And the total number of genes is 19,717.

Samtools and bedtools were applied to count number of reads for each base of the gene. Affected mapping reads were taken as the scale factor to normalize the reads for each individual. Hg19 human genome was taken as the reference.

Illumina paired-end RNA sequencing reads were aligned to GRCh37-lite genome-plus-junctions reference using BWA version 0.5.7. This reference combined genomic sequences in the GRCh37-lite assembly and exon-exon junction sequences whose corresponding coordinates were defined based on annotations of any transcripts in Ensembl (v59), Refseq, and known genes from the UCSC genome browser, which was downloaded on August 19, 2010, August 8, 2010, and August 19, 2010, respectively. Reads mapped to junction regions were then repositioned back to the genome and were marked with "ZJ:Z" tags. BWA is run using default parameters, except that the option (-s) is included to disable Smith-Waterman alignment. Finally, reads failing the Illumina chastity filter were flagged with a custom script, and duplicated reads were flagged with Picard's MarkDuplicates.

In order to make the data comparable, we applied log transformation on the observed expression profiles. Some genomic position has zero counts and we intentionally add 1 to it and then it returns to be zero after log transformation. After that expression counts for most of genes are of the same scale. We also mapped the genes onto the interval [0, 100].

3.2. Evaluation of the ODE for Modeling RNA-seq Data. To evaluate the precision of the ODE for modeling the RNA-seq data, we first used the ODE to fit the RNA-seq data where the coefficient functions were estimated. Then, we used numerical collocation Runge-Kutta method to solve the fitted ODE. The solutions of the fitted ODE as a function of genomic position were then compared with the observed RNA-seq curves.

We estimated the varying-coefficient functions using the proposed model. The expression function for gene $X(t)$ was first estimated by spline smoothing with some initial penalty. We then update the penalty using the proposed second order ODE with varying-coefficient functions. We iterated between curve smoothing and ODE estimation until convergence was achieved. The smoothing parameters $\boldsymbol{\lambda}$ were chosen by cross validation process. By selecting the value of $\boldsymbol{\lambda}$, we trade off basis expansion fitting error and ODE solution filtering error.

Larger value of $\boldsymbol{\lambda}$ put more emphasis on the ODE penalty and the solution to ODE with estimated parameters is more likely to approximate the original data.

To validate the estimates of coefficients functions in the model, it is essential to compare the observed gene expression curve to the ODE solution with estimated coefficient functions. We solved ODE numerically by considering two-point boundary value problems where boundary conditions are specified at both ends of the range of integration. We estimated two initial values at both ends by evaluating the estimated smoothing expression curves at start and end positions.

Figures 1(a) and 1(b) are fitted results of normal sample and cancer sample, respectively, for gene *CD74*. In these figures, the circles represent observed RNA-seq expression signal (green: normal; red: cancer); the blues lines are Fourier basis expansion to approximate the observed signal using weighted least square methods. The numbers of basis are chosen based on the length of genes and experimental adjustment to capture the important characteristics of gene expression curves. The dashed lines are estimated ODE solution using boundary value problem solver (R package *bvpSolve*). The ODE solutions approximate well the observed expression level of gene *CD74*, which show that estimated coefficient functions carry essential information of the original data. Once we have them, we can retrieve the original data very well.

To further evaluate the precision of the ODE for modeling RNA-seq data, we perform 5-fold cross validation prediction for gene *RPL29*. This method uses part of the available data to fit the model and estimate the parameters and uses the remaining data to test the model validity and estimate accuracy. We randomly split normal and cancer samples into five groups. From the estimation of parameters in the training samples, we solved the ODE with estimated coefficient functions to predict the expression curves of test samples. To be consistent, we estimated two initial values at both ends by evaluating the estimated smoothing expression curves at start and end positions in test samples. We also calculated the root mean square prediction error (RMSPE) for each folder to evaluate the performance of the prediction which is defined by

$$\text{RMSPE}_j = \frac{1}{N_j}\sum_{j=1}^{N_j}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{y}_i - y_i\right)^2}, \tag{35}$$

where y_i and $\widehat{y}_i$ are observed and predicted expression level; N is the number of genomic positions where the RNA-seq is observed for the gene and N_j is the number of subjects in the folder j.

Table 1 lists RMSPE in each folder for normal and cancer groups. The normal group has slightly better performance in terms of prediction on the test samples. But both prediction errors are relatively small.

Figures 2(a) and 2(b) are prediction results for selected samples in test set for gene *RPL29* in normal and cancer group, respectively. The gray dot is observed expression profile, and the solid red lines are Fourier basis expansion approximations to the observed expression data. The dashed green lines indicate the predicted gene expression profile in the test set by solving the estimated ODE in the training examples. We can observe that all of the prediction can capture the overall shape and fluctuation in the data. Secondly, they can also predict the magnitude of expression value with decent accuracy. These are predicted very close to the observed expression profiles.

Table 1: RMSPE in each folder for normal and cancer groups.

Folder list	Normal RMSPE	Cancer RMSPE
1	0.23	0.97
2	0.31	0.94
3	0.24	0.79
4	0.33	0.70
5	0.30	0.73

Table 2: The average sensitivity, specificity, and accuracy of top 12 genes to classify normal and KIRC group over 5-fold cross validation.

Genes	Sensitivity	Specificity	Accuracy
RBBP8	0.903	0.958	0.931
ZFYVE16	0.903	0.958	0.931
LOC100129034	0.889	0.944	0.917
SLC44A2	0.931	0.903	0.917
TTC21B	0.903	0.931	0.917
C18orf56	0.958	0.861	0.910
KCNJ16	0.889	0.931	0.910
PFKP	0.917	0.903	0.910
TMCC1	0.903	0.903	0.903
CDK18	0.917	0.875	0.896
SEC61G	0.903	0.889	0.896
ST6GAL1	0.861	0.931	0.896

3.3. Classification Analysis. These data suggest that the estimated coefficient functions capture important features of expression curves. From the solution to estimated ODE, we can see the exceptional retrieval of original data. From the prediction performance in the test set, we can also get well-predicted curves by just proving two initial boundary data points. It is natural to consider them as features to classify phenotype categories.

We obtain two coefficient functions from one expression function. We can use FPCA to help us to reduce the dimension of features and to ease the computational effort. We first applied FPCA technique on two coefficient functions $w_0(t)$ and $w_1(t)$ separately; then we combined two groups of the selected functional principal component scores as aggregated features before we provided them to classifier. In the end, we applied Lasso logistic regression to help us select features and make prediction on the groups.

Table 2 lists top 12 genes to differentiate normal and KIRC group using 5-fold cross validations. We can see that using a single gene it can reach as high as 90% classification accuracy. These data strongly indicate that the ODE model

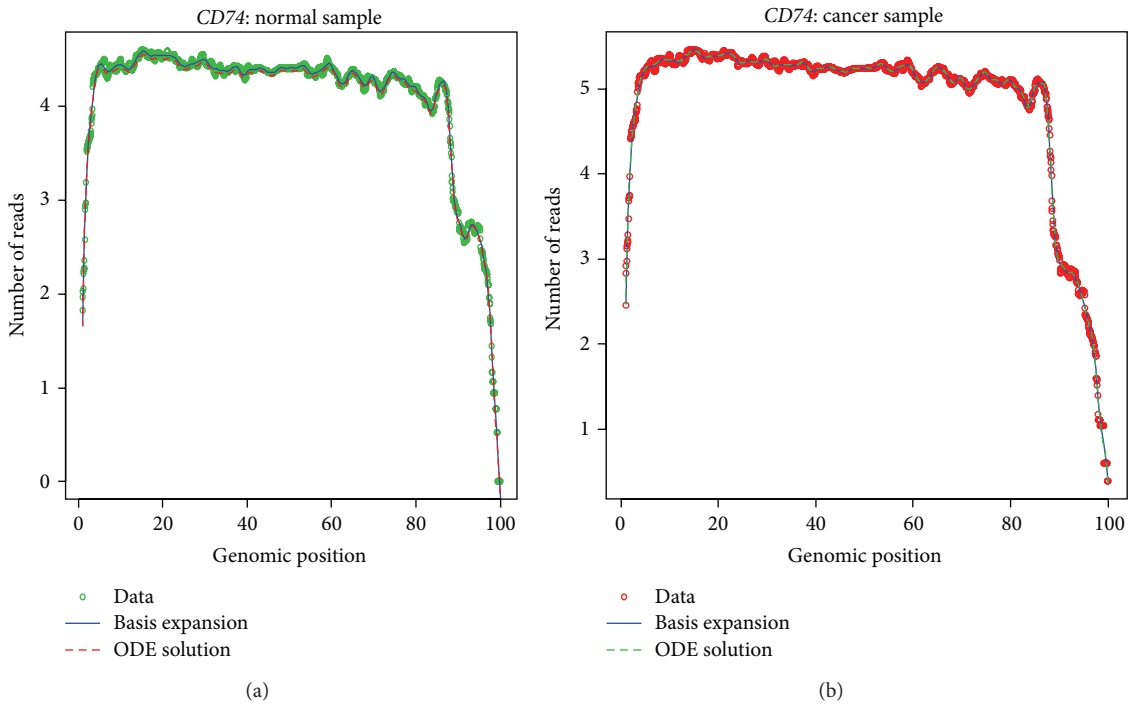


Figure 1: (a) Estimate of expression profiles for *CD74* by the ODE in a randomly selected normal sample. The green dotted points were observed expression levels, the blue solid lines are Fourier basis expansions, and the red dashed lines are numerical solution of ODE model. (b) Estimate of expression profiles for *CD74* by the ODE in a randomly selected tumor sample. The red dotted points were observed expression levels, the blue solid lines are Fourier basis expansions, and the green dashed lines are numerical solution of ODE model.

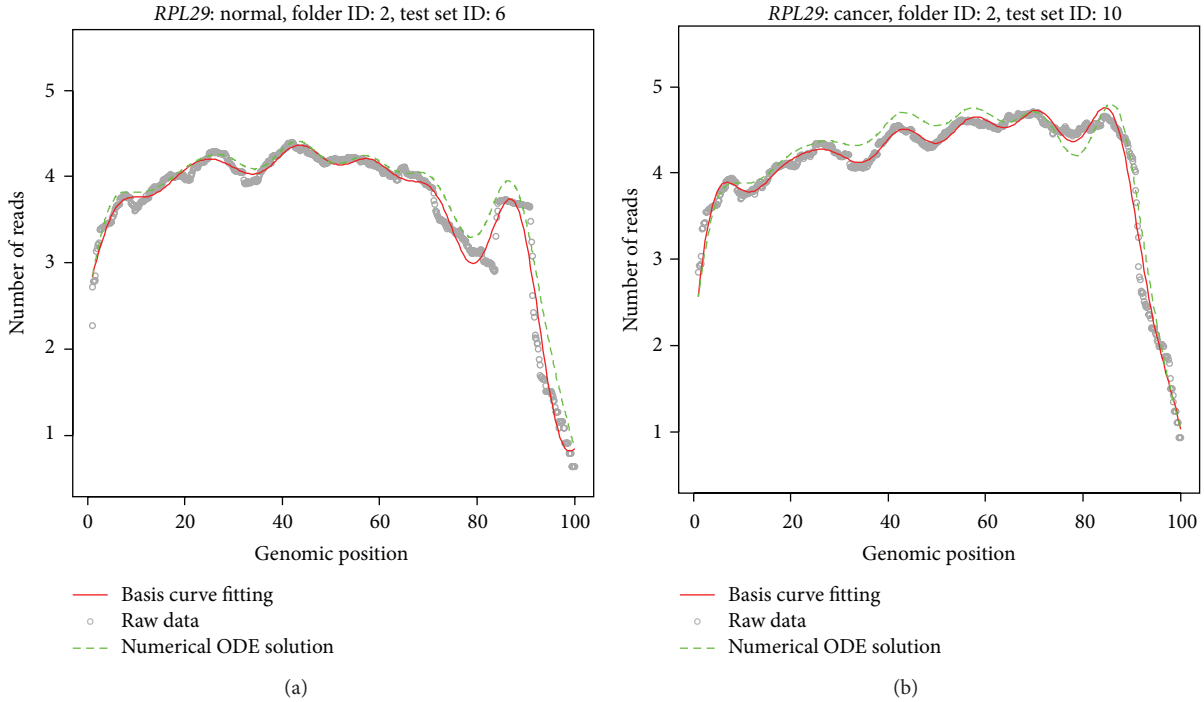


Figure 2: (a) Predicted expression curves for normal tissues for gene *RPL29*: The gray dot is observed expression profile, and the solid red lines are Fourier basis expansion approximation to the observed expression data. The dashed green lines are predicted gene expression profile in the test set by solving the estimated ODE in the training examples. (b) Predicted expression curves for tumor tissues for gene *RPL29*: The gray dot is observed expression profile, and the solid red lines are Fourier basis expansion approximation to the observed expression data. The dashed green lines are predicted gene expression profile in the test set by solving the estimated ODE in the training examples.

effectively captured the inherent features of RNA-seq expression profile. We also evaluated the performance of classification result using sensitivity, specificity, and accuracy. Sensitivity is defined as the percentage of cancer tissues correctly classified as cancer. Specificity is defined as the percentage of the normal tissues correctly classified as normal. The classification accuracy is defined as the percentage of the correctly classified normal and cancer tissues. The classification results can reach as high as 99% if we use all these 12 genes together as predictors.

3.4. Genome-Scale Clustering Analysis. In this section, we continue to use the estimated coefficient functions as features to cluster genes expression data to study the genome-wide transcriptome. By grouping genes with similar patterns of expression profiles, cluster analysis can provide insight into gene functions and biological process. It also gives a simple way of determining the functions of many genes for which information is not available, as genes with the same functions may share expression profiles. We assume the coefficient functions in ODE model help to define these patterns in the dynamic regulation process and give us clues to functional discovery and pattern grouping.

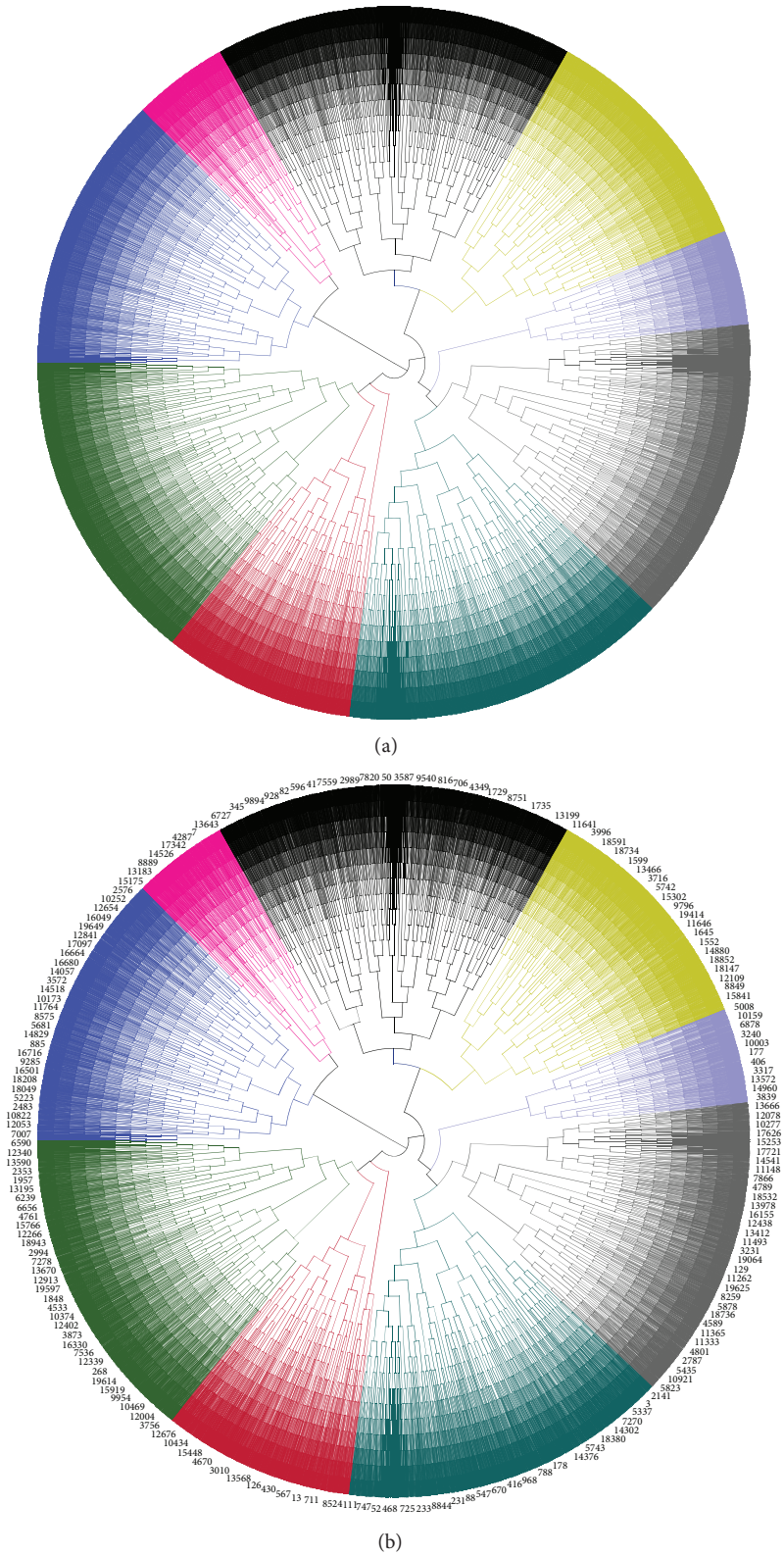
After we derive the feature matrix for all the genes from dimension reduction using FPCA, we merely need to adopt a metric definition which is used as a measure of similarity in the behavior of two genes. To calculate the distance matrix we used Euclidean distance and correlation matrix. This method computes a dendrogram that combines all genes in a single tree.

A total of 19717 genes were clustered into 9 groups according to the cluster analysis (Figures 3(a) and 3(b)). The functional principal component scores from coefficient functions in ODE model were used as significant features to define these patterns in the dynamic regulation process. The function annotation for each cluster was as follows.

The principle functions of the genes in the first group are mainly associated with oxidoreductase activity, ligase activity, dehydrogenase (NAD) activity, and related metabolic process. The detailed functions include aldehyde dehydrogenase (NAD) activity, translational initiation, mediator complex, MHC protein complex, mitochondrial membrane part, ion transmembrane transport, respiratory chain complex I, proton-transporting ATP synthase complex, proton-transporting two-sector ATPase complex, proton-transporting domain, NADH dehydrogenase (quinone) activity, oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor, positive regulation of protein ubiquitination, response to unfolded protein, heterocycle metabolic process, protein modification process, mitochondrial ATP synthesis coupled proton transport, glycerolipid metabolic process, macromolecule modification, proton-transporting two-sector ATPase complex, catalytic domain, regulation of translational initiation, oxidoreductase activity, acting on the aldehyde or oxo group of donors, RNA polymerase II transcription mediator activity, heme binding, positive regulation of ligase activity, negative regulation of ligase activity, negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle, positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle, regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle, positive regulation of ubiquitin-protein ligase activity, negative regulation of ubiquitin-protein ligase activity, transferase activity, glycerolipid biosynthetic process, amine transport, phosphoinositide metabolic process, carboxylic acid transport, hormone binding, eukaryotic translation initiation factor 3 complex, glycerophospholipid metabolic process, helicase activity, response to protein stimulus, lipid biosynthetic process, phosphoinositide biosynthetic process, aldehyde dehydrogenase [NAD(P)+] activity, proton-transporting ATP synthase complex, coupling factor F(o), cytosolic part, nucleobase, nucleoside and nucleotide metabolic process, proton-transporting ATPase activity, rotational mechanism, phospholipid metabolic process, phosphorus metabolic process, phosphate metabolic process, hydrogen-exporting ATPase activity, phosphorylative mechanism, proton-transporting V-type ATPase complex, MHC class II protein complex, collagen, positive regulation of protein modification process, and posttranslational protein modification.

The principle functions of the genes in the second group are mainly associated with hydratase activity, cation transmembrane transporter activity and hydrolase activity, and related metabolic process. The detailed functions include NAD or NADH binding, peroxisomal membrane, microbody membrane, aconitate hydratase activity, 4 iron, 4 sulfur cluster binding, regulation of vesicle-mediated transport, lactate dehydrogenase activity, L-lactate dehydrogenase activity, long-chain fatty acid-CoA ligase activity, fatty acid ligase activity, homophilic cell adhesion, tight junction, cation transmembrane transporter activity, occluding junction, kinesin complex, microbody part, peroxisomal part, actin filament binding, hydrolase activity, hydrolyzing O-glycosyl compounds, hydrolase activity, and acting on glycosyl bonds.

The principle functions of the genes in the third group are mainly associated with monooxygenase activity, receptor activity, electron carrier activity, sodium ion transmembrane transporter activity, and related metabolic process The detailed functions include nucleoside binding, purine nucleoside binding, monooxygenase activity, receptor activity, protein binding, ATP-binding, DNA packaging, chromatin assembly or disassembly, electron carrier activity, sodium ion transmembrane transporter activity, actin cytoskeleton, RNA metabolic process, adenyl nucleotide binding, GTPase regulator activity, regulation of lipid transport, negative regulation of lipid transport, adenyl ribonucleotide binding, protein-DNA complex, very-low-density lipoprotein particle, triglyceride-rich lipoprotein particle, cellular nitrogen compound metabolic process, cellular macromolecule biosynthetic process, nucleosome organization, chylomicron, organelle, intracellular organelle, cellular biosynthetic process, keratin filament, regulation of transcription, regulation of biological process, regulation of cellular process, regulation of nitrogen compound metabolic process, regulation of RNA metabolic process, regulation of macromolecule metabolic process, nucleoside-triphosphatase regulator activity, biological regulation, DNA conformation change, and regulation of primary metabolic process.



(a)

(b)

Figure 3: (a) Circular phylogram tree of 19717 genes that were clustered into nine groups by Dendroscope 3.2.10. (b) Detailed circular phylogram tree of 19717 genes that were clustered into nine groups by Dendroscope 3.2.10.

The principle functions of the genes in the fourth group are mainly associated with acyl-CoA thioesterase activity, oxidoreductase activity, phosphatase activity, and related metabolic process. The detailed functions include organellar small ribosomal subunit, organellar large ribosomal subunit, phospholipid-translocating ATPase activity, glutathione transferase activity, receptor signaling protein serine/threonine kinase activity, transmembrane receptor activity, inward rectifier potassium channel activity, organic acid transmembrane transporter activity, mitochondrial matrix, mitochondrial large ribosomal subunit, mitochondrial small ribosomal subunit, cytosol, translation, translational elongation, cell surface receptor linked signaling pathway, large ribosomal subunit, small ribosomal subunit, integral to membrane, acyl-CoA thioesterase activity, oxidoreductase activity, acting on NADH or NADPH, phosphatase activity, cytosolic ribosome, signaling process, signal transmission, intrinsic to membrane, negative regulation of protein ubiquitination, cullin-RING ubiquitin ligase complex, and mitochondrial lumen.

The principle functions of the genes in the fifth group are mainly associated with cell projection part, microtubule associated complex, motor activity, microtubule, axoneme, microtubule-based process, microtubule-based movement, microtubule cytoskeleton, dynein complex, cytoskeletal part, cilium, macromolecular complex, cilium axoneme, cell projection, protein complex, cilium part, pyrophosphatase activity, hydrolase activity, acting on acid anhydrides, hydrolase activity, acting on acid anhydrides, phosphorus-containing anhydrides, and nucleoside-triphosphatase activity.

The principle functions of the genes in the sixth group are mainly associated with intracellular signal transduction, cholesterol efflux, UDP-galactosyltransferase activity, and histone demethylase activity.

The principle functions of the genes in the seventh group are mainly associated with ATP-binding cassette (ABC) transporter complex, JNK cascade, ATP-dependent peptidase activity.

The principle functions of the genes in the eighth group are mainly associated with glutamate receptor activity, ATPase activity, cytoskeletal protein binding, and myosin filament.

The principle functions of the genes in the ninth group are mainly associated with adrenoceptor activity, inhibition of adenylate cyclase activity by G-protein signaling pathway, adenosine deaminase activity, hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, cyclic amidines, deaminase activity, adenylate cyclase activity, activation of protein kinase A activity, alpha-adrenergic receptor activity, adrenergic receptor binding, epinephrine binding, regulation of norepinephrine secretion, norepinephrine transport, positive regulation of blood pressure, norepinephrine secretion, oxidoreductase activity, acting on CH-OH group of donors, oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor, and delayed rectifier potassium channel activity.

Since the ODE model use the dynamic information of gene expression profiles and we consider genes with similar expression profiles will share common biological functions. Based on the cluster analysis, the genes grouped together have similar pattern of expression share common biological function. It directs us a way to find the functions of many genes for which information is unknown by looking the genes in the same group.

Table 3: Genes with significant difference in response behavior between normal and tumor samples.

ABHD10	*MFSD1*	*SDR39U1*	*ATP6V1D*	*OXA1L*	*SEC31A*
BST2	*PACSIN2*	*SMCR8*	*CD74*	*PGAM1P5*	*SSR2*
DAP3	*PIK3CB*	*TM9SF2*	*DHX40*	*PITRM1*	*UBXN6*
EDF1	*POLR2B*	*UQCRC2*	*HLA-DMB*	*PSAP*	*VKORC1*
HSPA9	*PSMB10*	*ZNF710*	*ISYNA1*	*MAT2A*	*PSMB7*
PSMC4					

3.5. Response Analysis of Gene Regulation. The expression level of a gene measured by sequencing can be viewed as a curve or function of genomic position. The gene expression will vary across the gene region. If we treat time and space position as the same argument, all theories and methods of dynamic system can be applied to RNA-seq data analysis. The dynamic behavior of a system is encoded in the temporal evolution of its states or in the genomic location evolution of the gene expression in our problem. Therefore, borrowing dynamic theory, we can study the location-dependent variation of gene expression under the perturbation of the external signals. The transient response of the dynamic systems is an important property of the system itself. It can be used to quantify the space domain characteristics of the gene regulation system responding to the disturbance of environments. Our goal is to investigate how the gene expression level at each genomic position varies in response to the external perturbation and whether this will affect the function of cell.

We conducted response analysis of 19,717 genes under unit-step signal perturbation. We used the Hoteling T^2 statistic that was described in Section 2.5 to identify 31 genes that showed significant difference in the response property. The names of 31 genes with significant difference in response property were summarized in Table 3. In a few cases, the matrix Λ may be singular; we can use penalized method or generalized inverse to estimate Λ^{-1}. However, this will inflate the false positive rates.

We present Figures 4(a)–4(d) showing the average expression curves, unit-step response curves, and the coefficient curves of the ODE of gene *CD74*, respectively. We observed that gene *CD74* not only showed significant difference in gene expression and coefficient curves of the ODE but also demonstrated strong difference in the unit-step response. The changing point of gene expression curve and unit-step response curve occurred between 11b and 12a where a splicing site is located. It was reported that CD74 played critical role in cancer cell tumorigenesis [16] and downregulation of CD74 inhibits growth and invasion in clear cell renal cell carcinoma [17].

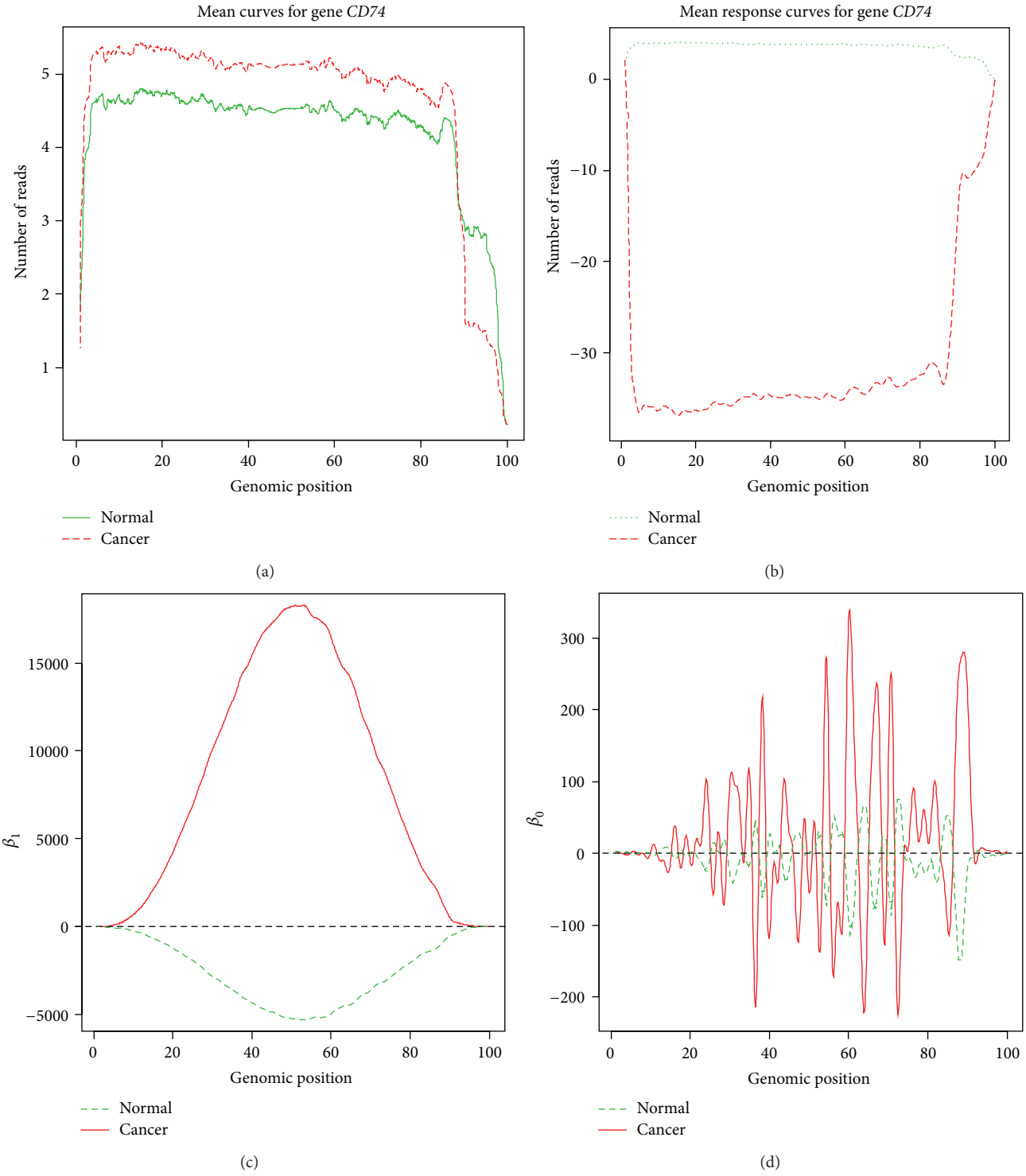


FIGURE 4: (a) Average expression curves of gene *CD74* in normal and tumor samples. (b) Average unit-step response curves of gene *CD74* in normal and tumor samples. (c) Average coefficient curves of the ODE for gene *CD74*. (d) Average coefficient curves of the ODE for gene *CD74*.

Transient response is one of dynamic properties. As shown in Figures S1A–D in Supplementary Material available online at http://dx.doi.org/10.1155/2015/916352, gene *ABHD10* that did not show significant difference in gene expression and coefficient curves of the ODE demonstrated strong difference in the unit-step response.

Figures S2A–D plotted the average expression curves, unit-step response curves, and the coefficient curves of the ODE of gene *BTS2*, respectively. Gene *BTS2* was differentially expressed but did not show significant difference in coefficient of the ODE between tumor and normal samples. Gene *BTS2* was identified to have significant difference in the unit-step response. The pattern of difference in the unit-step response may mainly due to rapid changes of gene expression in the region close to genomic position 20. From the literature we found that *BTS2* was associated with a number of cancers [18, 19].

4. Discussion

Dominant methods in literature for RNA-seq data analysis use a single valued summary statistic to represent expression level of a gene. However, a single number oversimplifies complex expression variation pattern across a gene and ignores information on alternative splicing and isoform and expression level variation at the genomic position level. To extract biologically useful expression variation signals across gene from RNA-seq data is a challenge, but important task. To meet this challenge, we have proposed using the ODE for modeling the RNA-seq data and addressed several essential issues for application of the ODE model to RNA-seq data analysis.

The first issue is how to use the ODE for modeling the RNA-seq data. We considered the number of reads or expression level at each position as a function of the genomic position and viewed the transcription process as a stochastic process of transcription along the gene. Borrowing dynamic theory from engineering, we have used the second order ODE to model the expression function of the gene measured by RNA-seq. We have employed differential principal analysis to develop statistical methods for estimation of location-varying coefficients of the ODE. We observed that the second order ODE almost has as good accuracy to predict gene expression as the third order ODE. But the third order ODE requires one more degree to describe the model. Therefore, the second order ODE is good enough to model gene expression.

The second issue is the precision of the ODE to model the RNA-seq data. We randomly split normal and cancer samples into five groups. From the estimation of parameters in the training samples, we solved the ODE with estimated coefficient functions to predict the expression curves of test samples. We have showed that the accuracy of the prediction by the second order ODE was very high and the root mean square prediction errors were quite small.

The third issue is how to extract useful regulatory signals from the RNA-seq data confound with measurement errors and sequencing technology variation. Since the second order ODE can model RNA-seq data very well, the location-coefficient functions of the ODE may well characterize the features of the regulatory process and measure the impact of the gene expression on the function of the cells and tissues. We have demonstrated that using location-coefficient functions of the second order ODE as features we have accurately classified the tumor and normal samples.

The fourth issue is to explore the applications of the ODE for RNA-seq data analysis. We have showed that the ODE can be used as a powerful tool to study the response of the gene transcription to the perturbation of environments. We have identified a number of cancer associated genes which showed significant difference in the response of the gene transcription between tumor and normal tissues but were not differentially expressed.

To our knowledge, this is the first time to use the ODE for modeling the RNA-seq data and investigation of gene transcription process. Our results were very preliminary. The samples were used to validate the accuracy of the ODE model to fit the real RNA-seq data. Large-scale validation and experiments for evaluating the model precision are urgently needed. Although the response analysis of dynamic model for the transcription process can help us to study how the external signals affect the gene expression variation across the gene, the mechanism of the gene transcription variation under the perturbation of external signals is largely unknown. The experiments for validation of the results of the response analysis of the dynamic models need to be performed. We lack consensus methods for RNA-seq data analysis. We are facing great challenges in developing innovative approaches and general framework for RNA-seq data analysis.

5. Conclusions

In conclusion, this study proposes the second order ODE for modeling RNA-seq data. We have demonstrated that the estimated ODE can accurately predict the gene expression level across the gene. We have showed that the location-dependent coefficients of the ODE effectively extract regulatory signals from the RNA-seq confounded with the measurement errors and sequencing technology variation and capture the inherent features of the transcription process. The results have showed that using coefficients of the ODE as features we can reach very high accuracy for classifying tumor and normal samples. Finally, we have demonstrated that using transient response analysis of dynamic system we identified 31 genes with significant differential response behavior between tumor and normal samples related to cancer.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The project described was supported by Grants 1R01AR057120-01 and 1R01HL106034-01 from the National Institutes of Health and NHLBI. The authors wish to thank TCGA working group for providing RNA-seq data.

The authors wish to acknowledge the contributions of the research institutions, study investigators, field staff, and study participants in creating the TCGA datasets for biomedical research.

References

[1] H. Xiong, J. B. Brown, N. Boley, P. J. Bickel, and H. Huang, "DE-FPCA: testing gene differential expression and exon usage through functional principal component analysis," in *Statistical Analysis of Next Generation Sequencing Data*, pp. 129–143, Springer, New York, NY, USA, 2014.

[2] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, 2011.

[3] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[4] F. Rapaport, R. Khanin, Y. Liang et al., "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biology*, vol. 14, no. 9, article R95, 2013.

[5] C. Suo, S. Calza, A. Salim, and Y. Pawitan, "Joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-Seq data," *Bioinformatics*, vol. 30, no. 4, Article ID btt704, pp. 506–513, 2014.

[6] Z. Sun and Y. Zhu, "Systematic comparison of RNA-Seq normalization methods using measurement error models," *Bioinformatics*, vol. 28, no. 20, pp. 2584–2591, 2012.

[7] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Research*, vol. 38, no. 12, article e131, 2010.

[8] K. D. Hansen, Z. Wu, R. A. Irizarry, and J. T. Leek, "Sequencing technology does not eliminate biological variability," *Nature Biotechnology*, vol. 29, no. 7, pp. 572–573, 2011.

[9] K. Ogata, *System Dynamics*, Prentice Hall, 3rd edition, 1997.

[10] L. Luo, E. Boerwinkle, and M. Xiong, "Association studies for next-generation sequencing," *Genome Research*, vol. 21, no. 7, pp. 1099–1108, 2011.

[11] J. O. Ramsay, *Functional Data Analysis*, Wiley Online Library, 2006.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, 2nd edition, 2009.

[13] K. Koh, S. J. Kim, and S. P. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.

[14] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[15] K. Soetaert, J. Cash, and F. Mazzia, *Solving Differential Equations in R*, Use R!, Springer, 2012.

[16] Y.-H. Liu, C.-Y. Lin, W.-C. Lin, S.-W. Tang, M.-K. Lai, and J.-Y. Lin, "Up-regulation of vascular endothelial growth factor-D expression in clear cell renal cell carcinoma by CD74: a critical role in cancer cell tumorigenesis," *The Journal of Immunology*, vol. 181, no. 9, pp. 6584–6594, 2008.

[17] S.-Q. Ji, X.-L. Su, W.-L. Cheng, H.-J. Zhang, Y.-Q. Zhao, and Z.-X. Han, "Down-regulation of CD74 inhibits growth and invasion in clear cell renal cell carcinoma through HIF-1α pathway," *Urologic Oncology: Seminars and Original Investigations*, vol. 32, no. 2, pp. 153–161, 2014.

[18] K. H. Fang, H. K. Kao, L. M. Chi et al., "Overexpression of BST2 is associated with nodal metastasis and poorer prognosis in oral cavity cancer," *The Laryngoscope*, vol. 124, no. 9, pp. E354–E360, 2014.

[19] A. Sayeed, G. Luciani-Torres, Z. Meng, J. L. Bennington, D. H. Moore, and S. H. Dairkee, "Aberrant regulation of the BST2 (Tetherin) promoter enhances cell proliferation and apoptosis evasion in high grade breast cancer cells," *PLoS ONE*, vol. 8, no. 6, Article ID e67191, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 461593, 10 pages
http://dx.doi.org/10.1155/2015/461593

Research Article

Robust Association Tests for the Replication of Genome-Wide Association Studies

Jungnam Joo,[1] Ju-Hyun Park,[2] Bora Lee,[1] Boram Park,[1] Sohee Kim,[1] Kyong-Ah Yoon,[3] Jin Soo Lee,[3] and Nancy L. Geller[4]

[1] *Biometric Research Branch, Research Institute and Hospital, National Cancer Center, Gyeonggi-do, Goyang-si 410-769, Republic of Korea*
[2] *Department of Statistics, Dongguk University, Seoul 100-715, Republic of Korea*
[3] *Lung Cancer Branch, Research Institute and Hospital, National Cancer Center, Gyeonggi-do, Goyang-si 410-769, Republic of Korea*
[4] *Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892-7938, USA*

Correspondence should be addressed to Jungnam Joo; jooj@ncc.re.kr

Received 14 November 2014; Revised 14 February 2015; Accepted 14 February 2015

Academic Editor: Taesung Park

In genome-wide association study (GWAS), robust genetic association tests such as maximum of three CATTs (MAX3), each corresponding to recessive, additive, and dominant genetic models, the minimum p value of Pearson's Chi-square test with 2 degrees of freedom, and CATT based on additive genetic model (MIN2), genetic model selection (GMS), and genetic model exclusion (GME) methods have been shown to provide better power performance under wide range of underlying genetic models. In this paper, we demonstrate how these robust tests can be applied to the replication study of GWAS and how the overall statistical significance can be evaluated using the combined test formed by p values of the discovery and replication studies.

1. Introduction

With the advance of biotechnology and substantial reduction of genotyping costs, a genome-wide association study (GWAS) using hundred thousand markers in several thousand individuals is now increasingly utilized and has been successful in detecting genetic associations across the entire genome with complex human traits [1–6]. Among many challenges this application holds; development of more efficient and robust statistical methodologies with higher power to detect an association with a single marker has been one of the most important statistical issues, given that effects of individual markers are usually characterized as being small to moderate. One attempt to overcome this challenge is focused on developing efficient tests that are robust against underlying genetic model misspecification.

Two most frequently used association tests are the allele-based test (ABT) and the genotype-based test (GBT). ABT compares the allele frequencies between cases and controls, while GBT compares the genotype distributions of cases and controls. The Cochran-Armitage trend test (CATT) [7, 8] is a popular GBT which takes into account the underlying genetic model. It is well known, however, that the ABT may inflate type I error when Hardy-Weinberg equilibrium (HWE) does not hold in the samples [9]. Even under HWE, when the genetic model is recessive or dominant, the ABT may suffer from serious power loss. On the other hand, the CATT does not depend on HWE, but to apply the CATT the choice of scores optimal for the underlying genetic model needs to be specified. For complex diseases, the genetic model is usually unknown and robust tests such as the maximum of three CATTs (MAX3) [10] and the maximum efficiency robust test (MERT) [11, 12] are preferable. Alternatively, Zheng and Ng [13] and Joo et al. [14] proposed a two-phase analysis based on the genetic model selection (GMS) and genetic model exclusion (GME). Moreover, an alternative approach was proposed by the Wellcome trust case-control consortium (WTCCC) [5] which used a minimum p value of Pearson's Chi-square test and additive CATT, and the asymptotic properties of this approach were studied in detail by Joo et al. [15]. These methods provide better or comparable power performance than some of the robust tests such as MAX3.

In this paper, we illustrate how these robust tests can be applied to a replication study of GWAS and how overall statistical significance can be evaluated using the combined test formed by p values of the discovery and replication studies. The importance of replication or validation in GWAS has been well recognized [16, 17], and joint analysis in a two-stage design of GWAS has been proved to be more powerful than replication-based analysis and has been widely conducted in GWAS with a variety of phenotypes of interest [18, 19].

The paper is organized as follows. We first describe the data structures and notation and review existing robust association tests for a single data set. Then we describe how to obtain the p value for the replication data set, given the significant result of the discovery stage, using robust tests. In the next section, a combined test of the p values of the discovery and replication data sets is proposed, together with the way to evaluate the statistical significance for the combined test. Simulation studies are conducted to compare the type I error rates and powers of various analytical strategies. For illustration purposes, the summarized methods are applied to a non-small-cell lung cancer data set and at the end there is a discussion.

2. Methods

2.1. Data and Notation. For a marker with two alleles A and B, let the frequencies of B in cases and controls be $p = P(B \mid \text{case})$ and $q = P(B \mid \text{control})$. Denote three genotypes by $G_0 = AA$, $G_1 = AB$, and $G_2 = BB$. In case-control association studies, r cases and s controls are independently sampled from each population. The observed genotype counts for (G_0, G_1, G_2) are (r_0, r_1, r_2) in the cases and (s_0, s_1, s_2) in the controls. Disease prevalence is denoted by $k = P(\text{disease})$ and penetrance by $f_i = P(\text{disease} \mid G_i)$ for $i = 0, 1, 2$. Two genotype relative risks (GRRs) are denoted by $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$ using $f_0 > 0$ as baseline penetrance. Under the null hypothesis of no association $H_0 : f_0 = f_1 = f_2 = k$ or alternatively $H_0 : \lambda_2 = \lambda_1 = 1$. Genetic model is recessive (REC), additive (ADD), multiplicative (MUL), and dominant (DOM) when $\lambda_1 = 1$, $\lambda_1 = (1+\lambda_2)/2$, $\lambda_1 = \lambda_2^{1/2}$, and $\lambda_2 = \lambda_1$, respectively.

2.2. Review of Association Tests for a Single Data Set. The association in case-control studies can be tested using various methods which have been extensively studied. The general association between the disease status and the SNP can be tested using Pearson's Chi-square test which has an asymptotic Chi-square distribution with 2 degrees of freedom under H_0. The test is given by

$$T_{\text{chi2}} = \sum_{j=0}^{2} \frac{\left(r_j - n_j r/n\right)^2}{n_j r/n} + \sum_{j=0}^{2} \frac{\left(s_j - n_j s/n\right)^2}{n_j s/n}, \tag{1}$$

where $n_i = r_i + s_i$ for $i = 0, 1, 2$ and $n = r + s$. Under Hardy-Weinberg equilibrium (HWE), an allele-based test (ABT) and CATT with scores $(0, x, 1)$ for (G_0, G_1, G_2), where $0 \le x \le 1$, are given by

$$\begin{aligned} Z_{\text{ABT}} &= \frac{n^{1/2}\left\{2r\left(2s_0 + s_1\right) - 2s\left(2r_0 + r_1\right)\right\}}{\left\{2rs\left(2n_0 + n_1\right)\left(n_1 + 2n_2\right)\right\}^{1/2}}, \\ Z_x &= \frac{n^{1/2}\sum_{i=0}^{2} x_i\left(sr_i - rs_i\right)}{\left[rsn\left\{n\sum_{i=0}^{2} x_i^2 n_i - \left(\sum_{i=0}^{2} x_i n_i\right)^2\right\}\right]^{1/2}}, \end{aligned} \tag{2}$$

where $(x_0, x_1, x_2) = (0, x, 1)$ [9]. The optimal choices of x for the recessive (REC), additive/multiplicative (ADD/MUL), and dominant (DOM) models are $x = 0, 1/2$ and 1, respectively [9, 20]. Both Z_x and Z_{ABT} asymptotically follow a standard normal distribution under H_0. Z_x can be used even when HWE does not hold. However, without the HWE assumption, Z_{ABT} does not follow a standard normal distribution due to the correlation between two alleles.

A robust test, MAX3 proposed by Friedlin et al. [10], can be obtained by taking the maximum of three CATTs under the three genetic models as $\text{MAX3} = \max(|Z_0|, |Z_{1/2}|, |Z_1|)$. Parametric bootstrap or permutation methods can be used to find the p value of MAX3 [4].

Let the p values of Pearson's Chi-square test and CATT under the additive genetic model $Z_{1/2}$ be P_{chi2} and $P_{1/2}$, respectively. WTCCC [5] proposed an alternative robust test $\text{MIN2} = \min(P_{\text{chi2}}, P_{1/2})$. Joo et al. [15] derived the asymptotic null distribution of MIN2 and using their result the p value of MIN2 can be obtained as

$$\begin{aligned} P_{\text{MIN2}} &= \frac{1}{2}\exp\left\{-\frac{1}{2}H_1^{-1}\left(1 - \text{MIN2}\right)\right\} + \frac{1}{2}\text{MIN2} \\ &\quad - \frac{1}{2\pi}\int_{H_1^{-1}(1-\text{MIN2})}^{-2\log(\text{MIN2})} e^{-v/2}\arcsin\left(\frac{2H_1^{-1}\left(1 - \text{MIN2}\right)}{v - 1}\right)dv, \end{aligned} \tag{3}$$

where H_1 and H_2 are the cumulative distributions of Chi-square distributions with 1 and 2 degrees of freedom.

On the other hand, Song and Elston [21] considered a Hardy-Weinberg disequilibrium trend test (HWDTT) given by

$$Z_H = \frac{(rs/n)^{1/2}\left(\widehat{\Delta}_p - \widehat{\Delta}_q\right)}{\left\{1 - n_2/n - n_1/(2n)\right\}\left\{n_2/n + n_1/(2n)\right\}}, \tag{4}$$

where $\widehat{\Delta}_p = \widehat{p}_2 - (\widehat{p}_2 + \widehat{p}_1/2)^2$ and $\widehat{\Delta}_q = \widehat{q}_2 - (\widehat{q}_2 + \widehat{q}_1/2)^2$ are the estimates of Δ_p and Δ_q, where $\widehat{p}_i = r_i/r$ and $\widehat{q}_i = s_i/s$. Here, Δ denotes the Hardy-Weinberg disequilibrium (HWD) coefficient defined by $Pr(BB) - \{Pr(AB)/2 + Pr(BB)\}^2$ and Δ_p and Δ_q denote the HWD coefficient in cases and controls, respectively.

Zheng and Ng [13] used the information contained in the signs of (Δ_p, Δ_q) to determine the genetic models in their two-phase method. Their two-phase statistic Z_{GMS} is given by $Z_{\text{GMS}} = Z_0$ if $Z_H > c$, Z_1 if $Z_H < -c$, and $Z_{1/2}$ otherwise, where $c = \Phi^{-1}(1 - \alpha_H)$ for $\alpha_H = 0.05$. The asymptotic

correlations between Z_H and three CATTs under HWE were derived and the significance level was adjusted accordingly to control the desired type I error. Based on the observation that this method assumes B is the risk allele, Joo et al. [14] studied the behavior of Z_{GMS} when either one of the alleles can be a risk allele. They chose the risk allele based on the sign of $Z_{1/2}$; that is, if $Z_{1/2} > 0$, B is the risk allele, and Z_0, $Z_{1/2}$, and Z_1 are chosen for REC, ADD, and DOM models, respectively. If $Z_{1/2} < 0$, the respective test statistics are chosen to be $-Z_1$, $-Z_{1/2}$, and $-Z_0$. They incorporate this property in defining the test statistic for genetic model selection (Z_{GMS}) and calculating the p value. Let $\Theta_0(z) = \{z : z > c\}$, $\Theta_{1/2}(z) = \{z : |z| < c\}$, and $\Theta_1(z) = \{z : z < -c\}$. Then, the p value of this method can be obtained by

$$
\begin{aligned}
&p_{\mathrm{GMS}} \\
&= 2\left\{\sum_{x=0}^{1}\int_{\Theta_x(z_H)}\int_0^{\infty}\int_t^{\infty}\phi_x\left(z_x, z_{1/2}, z_H\right)dz_x dz_{1/2}dz_H\right\} \\
&\quad + 2\left\{\int_{\Theta_{1/2}(z)}\Phi\left(\frac{(-t\wedge 0)+\rho_{1/2}z}{\left(1-\rho_{1/2}^2\right)^{1/2}}\right)d\Phi\left(z\right)\right\},
\end{aligned}
\tag{5}
$$

where $\rho_x = \mathrm{Corr}(Z_x, Z_H)$ in (5) and $\rho_{x,1/2} = \mathrm{Corr}(Z_x, Z_{1/2})$ ($x = 0, 1$) are replaced by their consistent estimates. Here, $t = z_{\mathrm{GMS}}$ and $(-t \wedge 0) = \min(-t, 0)$. Moreover, z_{GMS} and $z_{1/2}$ are the observed values of Z_{GMS} and $Z_{1/2}$, respectively.

While studying the properties of GMS, Joo et al. [14] noticed that the probability of selecting the true recessive or dominant models using Z_H is very low especially for low to moderate GRRs, but the unlikely genetic model can be successfully excluded. This led to genetic model exclusion method Z_{GME} which is the same as the Z_{GMS} described above except Z_x for $x = 0, 1/2, 1$ is replaced by Z_x^* where $Z_x^* = (Z_x + Z_{1/2})/\{2(1+\widehat{\rho}_{x,1/2})\}^{1/2}$. And the p value of GME can be obtained as

$$
\begin{aligned}
&p_{\mathrm{GME}} \\
&= 2\left\{\sum_{x=0}^{1}\int_{\Theta_x(z_H)}\int_0^{\infty}\int_L^{\infty}\phi_x\left(z_x, z_{1/2}, z_H\right)dz_x dz_{1/2}dz_H\right\} \\
&\quad + 2\left\{\int_{\Theta_{1/2}(z)}\Phi\left(\frac{(-t\wedge 0)+\rho_{1/2}z}{\left(1-\rho_{1/2}^2\right)^{1/2}}\right)d\Phi\left(z\right)\right\},
\end{aligned}
\tag{6}
$$

where $L = t\{2(1+\widehat{\rho}_{x,1/2})\}^{1/2} - z_{1/2}$ for $t = z_{\mathrm{GME}}$.

2.3. p Value of Replication Data Using the Robust Method. In the discovery stage, the p value of robust association tests, including MAX3, MIN2, Z_{GMS}, and Z_{GME}, can be obtained as described in Section 2.2. For the p value of replication data using the robust method, we use the same analytic method that was used for discovery and the risk allele identified by it [16]. This means that when the best test statistic or genetic model is selected in the discovery stage, the replication stage will adopt the discovery stage selection and the direction of association.

Suppose that, for simplicity of notation, our interest is in GWAS with two stages, one for discovery and the other for replication, although the methodology described below can be extended to multistages for replication. Let $Z_x^{(i)}$ for $x = 0, 1/2, 1$ be the CATT optimal for recessive, additive, and dominant models and let $P_x^{(i)}$ be corresponding p value for ith stage ($i = 1$ for discovery and $i = 2$ for replication stages). Also, denote $Z_x^{(i)*} = (Z_x^{(i)} + Z_{1/2}^{(i)})/\{2(1+\widehat{\rho}_{x,1/2}^{(i)})\}^{1/2}$ for $x = 0, 1/2, 1$. Then, for CATT with a preselected genetic model, $P_x^{(2)} = 1 - \Phi(\mathrm{sign}(Z_x^{(1)} \cdot Z_x^{(2)}) \cdot |Z_x^{(2)}|)$ using a one-sided p value given the direction of association from the discovery stage, and $P_x^{(2)*} = 1 - \Phi(\mathrm{sign}(Z_x^{(1)*} \cdot Z_x^{(2)*}) \cdot |Z_x^{(2)*}|)$. Moreover, denote the test statistics and p values using Pearson's Chi-square test from the ith stage as $T_{\mathrm{chi2}}^{(i)}$ and $P_{\mathrm{chi2}}^{(i)}$. Further, let HWDTT from the ith stage be $Z_H^{(i)}$. Then, the second stage p values, using MAX3, MIN2, Z_{GMS}, and Z_{GME}, denoted as $P_{\mathrm{MAX3}}^{(2)}$, $P_{\mathrm{MIN2}}^{(2)}$, $P_{\mathrm{GMS}}^{(2)}$, and $P_{\mathrm{GME}}^{(2)}$, can be obtained as follows:

$$
\begin{aligned}
P_{\mathrm{MAX3}}^{(2)} &= P_{x^*}^{(2)}, \quad \text{where } x^* = \arg\min_{x\in\{0,1/2,1\}} P_x^{(1)}, \\
P_{\mathrm{MIN2}}^{(2)} &= P_{1/2}^{(2)} \cdot I\left(P_{1/2}^{(1)} \le P_{\mathrm{chi2}}^{(2)}\right) + P_{\mathrm{chi2}}^{(2)} \cdot I\left(P_{1/2}^{(1)} > P_{\mathrm{chi2}}^{(2)}\right), \\
P_{\mathrm{GMS}}^{(2)} &= P_0^{(2)} I\left(Z_H^{(1)} > c\right) + P_1^{(2)} I\left(Z_H^{(1)} < -c\right) \\
&\quad + P_{1/2}^{(2)} I\left(\left|Z_H^{(1)}\right| \le c\right), \quad \text{if } Z_{1/2}^{(1)} > 0; \\
&= P_0^{(2)} I\left(Z_H^{(1)} < -c\right) + P_1^{(2)} I\left(Z_H^{(1)} > c\right) \\
&\quad + P_{1/2}^{(2)} I\left(\left|Z_H^{(1)}\right| \le c\right), \quad \text{if } Z_{1/2}^{(1)} \le 0, \\
P_{\mathrm{GME}}^{(2)} &= P_0^{(2)*} I\left(Z_H^{(1)} > c\right) + P_1^{(2)*} I\left(Z_H^{(1)} < -c\right) \\
&\quad + P_{1/2}^{(2)*} I\left(\left|Z_H^{(1)}\right| \le c\right), \quad \text{if } Z_{1/2}^{(1)} > 0; \\
&= P_0^{(2)*} I\left(Z_H^{(1)} < -c\right) + P_1^{(2)*} I\left(Z_H^{(1)} > c\right) \\
&\quad + P_{1/2}^{(2)*} I\left(\left|Z_H^{(1)}\right| \le c\right), \quad \text{if } Z_{1/2}^{(1)} \le 0.
\end{aligned}
\tag{7}
$$

It is important to note that even though the direction of the test statistics and the selected genetic models are used to obtain the second stage p values, the p values from the two stages are independent under the null hypothesis. This is because, under the null hypothesis, the probability of $Z_{1/2}$ being positive or negative is simply 1/2, and the probability of the selection of a certain genetic model is also a constant (α_H for the recessive and dominant models and $1 - 2\alpha_H$ for the additive model).

2.4. Combined Test Using p Values and Its Statistical Significance. For a given robust test, we can consider the joint analysis by combining p values from the discovery and replication stages of GWAS. We consider using p values rather than the test statistics because test statistics can have

Table 1: Type I error rates of three approaches—replication-based (REP) test, Fisher's combination (Z_{FC}), and linear combination of test (Z_{LC})—based on the CATT with an additive model ($Z_{1/2}$), χ^2, MAX3, MIN2, GMS, and GME. The disease prevalence $K = 0.1$, $M = 10$ markers, $r = 1,500$ cases, and $s = 1,500$ controls are considered based on 20,000 simulations.

MAF	π_s	α_D		$F = 0$					
				$Z_{1/2}$	χ^2	MAX3	MIN2	GMS	GME
0.3	0.5	0.05	REP	0.00530	0.00455	0.00505	0.00485	0.0050	0.00490
			Z_{FC}	0.00500	0.00535	0.00495	0.00510	0.00515	0.00460
			Z_{LC}	0.00535	0.00525	0.00485	0.00510	0.0050	0.00485
		0.1	REP	0.00510	0.00560	0.00565	0.00485	0.00525	0.00545
			Z_{FC}	0.00565	0.00535	0.00565	0.00545	0.00565	0.00540
			Z_{LC}	0.00520	0.00565	0.00525	0.00520	0.00530	0.00525
0.3	0.6	0.05	REP	0.00510	0.00485	0.00480	0.00515	0.00480	0.00500
			Z_{FC}	0.00445	0.00455	0.00450	0.00455	0.00450	0.00460
			LC	0.00500	0.00515	0.00495	0.00520	0.00475	0.00480
		0.1	REP	0.00500	0.00485	0.00485	0.00535	0.00530	0.00505
			Z_{FC}	0.00465	0.00490	0.00485	0.00500	0.00455	0.00460
			Z_{LC}	0.00480	0.00470	0.00515	0.00490	0.00485	0.00475
0.4	0.5	0.05	REP	0.00590	0.00505	0.00530	0.00565	0.00505	0.00510
			Z_{FC}	0.00575	0.00430	0.00460	0.00535	0.00460	0.00500
			Z_{LC}	0.00600	0.00445	0.00500	0.00540	0.00490	0.00490
		0.1	REP	0.00525	0.00470	0.00535	0.00450	0.00480	0.00515
			Z_{FC}	0.00515	0.00510	0.00495	0.00475	0.00540	0.00500
			Z_{LC}	0.00530	0.00500	0.00485	0.00475	0.00495	0.00510
0.4	0.6	0.05	REP	0.00475	0.00585	0.00480	0.00500	0.00515	0.00495
			Z_{FC}	0.00460	0.00470	0.00420	0.00490	0.00455	0.00440
			Z_{LC}	0.00525	0.00550	0.00520	0.00580	0.00510	0.00510
		0.1	REP	0.00550	0.00490	0.00515	0.00535	0.00555	0.00540
			Z_{FC}	0.00520	0.00370	0.00495	0.00450	0.00515	0.00510
			Z_{LC}	0.00565	0.00485	0.00570	0.00530	0.00610	0.00580

complex forms and obtaining the distribution of the joint test can be difficult. On the other hand, calculating a p value for each data set might be relatively simple, and the distribution of p values under the null hypothesis of no association is easy to handle.

There are several methods for combining test statistics from two stages [22], and two most commonly used forms are based on Fisher's combination and a linear combination after inverse normal transformation [23]. Fisher's combination (FC) directly sums p values after $-2\log$ transformation; that is, $Z_{FC} = -2w_1 \log(P^{(1)}) - 2w_2 \log(P^{(2)})$, where $P^{(i)}$ is p value from $i = 0$ for discovery and $i = 1$ for replication stages using a given robust test. A specification of $w_1 = w_2 = 1$ gives the same weight for discovery and replication stages, and one can consider $w_1 = 2\pi_s$ and $w_2 = 2(1 - \pi_s)$ where $\pi_s = N_D/(N_D + N_R)$, and N_D and N_R are sample sizes of the discovery and replication data sets. A linear combination of two P values after taking the inverse of the standard normal cumulative distribution is given by $Z_{LC} = \{w_1\Phi^{-1}(1-P^{(1)}/2)+w_2\Phi^{-1}(1-P^{(2)})\}/\sqrt{w_1^2 + w_2^2}$ with a natural choice of $w_1 = \sqrt{\pi_s}$ and $w_2 = \sqrt{1 - \pi_s}$. Let the significance level of the discovery stage be α_D, which means that markers with $P^{(1)} < \alpha_D$ are selected and replicated in the replication stage. The p value of combined test can then be obtained as $p_{FC} = P_{H_0}(P^{(1)} < \alpha_D, Z_{FC} > z_{FC})$ where the observed value of Z_{FC} is z_{FC}. The p_{FC} are calculated as $e^{-z_{FC}/2}(1+z_{FC}/2+\log\alpha_D)$ for equal weights where $z_{FC} > -2\log\alpha_D$ and $(w_1/(w_1 - w_2))e^{-z_{FC}/2w_1} - (w_2/(w_1-w_2))e^{-z_{FC}/2w_2}\alpha_d^{-(w_1-w_2)/w_2}$ for unequal weights where $z_{FC} > -2w_1\log\alpha_D$. Detailed derivations are described in the Appendix. Equivalently, for an overall type I error threshold for a single marker of α, one may obtain the threshold C_{FC} of Z_{FC} that satisfies $P_{H_0}(P^{(1)} < \alpha_D, Z_{FC} > C_{FC}) \leq \alpha$. Similarly, for the Z_{LC}, the p value is calculated as $p_{LC} = P_{H_0}(P^{(1)} < \alpha_D, Z_{LC} > z_{LC}) = \int_{z_{1-\alpha_D/2}}^{\infty} \phi(z)[1 - \Phi((\sqrt{w_1^2 + w_2^2}z_{LC} - w_1 z)/w_2)]dz$ for $z_{LC} > z_{1-\alpha_D/2}$ where the observed value of $Z_{LC} = z_{LC}$.

3. Simulation Results

3.1. Type I Error. Table 1 provides the type I errors under different scenarios. A disease prevalence of 10% is assumed, and a total of 1500 cases and 1500 controls were divided into two stages. The proportions of samples in the first stage (π_s) of 0.5 and 0.6 were considered for the minor allele frequency (MAF) of 0.3 and 0.4. We considered $M = 10$ markers to control the genome-wide false positive rate at $\alpha = 0.05$ with the Bonferroni correction. We did not consider a larger M

such as 300,000 or 500,000 because this will require more than 10 million simulations to show a stable estimate of the type I error rate. With $M = 10$, we performed 20,000 simulations which result in less than 10% of a coefficient of variation for a significance level $0.05/M = 0.005$ for each marker [24]. The test statistics considered are $Z_{1/2}$, Pearson's Chi-square test, MIN2, MAX3, GMS, and GME. For the second stage analysis, we considered a replication-based analysis, Z_{FC}, and Z_{LC} as proposed above. The results are based on the situation under HWE (HWE coefficient $F = 0$). As expected, all tests control the type I error reasonablly well, and similar results were obtained when a slight deviation from HWE is present with $F = 0.05$ (results not shown).

3.2. Empirical Power. We examined the empirical powers of different tests considered above. In Figure 1, we considered $M = 10$ markers, a disease prevalence of 10%, the same genotype relative risk for two stages ($r_1 = 1.4$ and $r_2 = 1.4$), and 1,000 cases and 1,000 controls. 2,000 simulations were performed under HWE ($F = 0$) to control the genome-wide false positive rate at $\alpha = 0.05$. The recessive, additive, and dominant models were assumed for the first, second, and third rows. Both joint analyses showed better power performances compared to the replication-based analysis (up to 15.9% in scenarios considered in Figure 1), and LC and FC have comparable powers with less than 2% difference. The power gain of using the joint analysis is not as much as that observed in Skol et al. [18]. However, as reported by Skol et al. [18], when the between-stage heterogeneity exists and the risk allele has a larger effect in the first stage than that in the second stage, much improved power is observed by using the joint test. Figure 2 shows results under this scenario with $r_1 = 1.6$ and $r_2 = 1.4$, and the observed increase in power using the joint test is as high as 33.9%. Again, the difference between LC and FC is minor with less than 3% difference. As for comparison between different robust methods, MAX3, GMS, and GME perform well under the recessive model, while $Z_{1/2}$, χ^2, and MIN2 are less powerful. Under the additive model, $Z_{1/2}$ is most powerful, as expected, and χ^2 is least powerful. Other robust methods perform well with a slight decrease in power compared to $Z_{1/2}$. Under the dominant model, MAX3, GMS, and GME perform the best even though all tests show good power performances, and the difference is minor. Similar patterns were observed when a slight deviation from the HWE is present (results not shown).

4. Real Data Application

The GWAS on non-small-cell lung cancer (NSCLC) by Yoon et al. [25] studied 621 NSCLC patients and 1541 control subjects in the discovery stage. After stringent quality control steps, a total of 246,758 SNPs were tested for the association with NSCLC based on $Z_{1/2}$. In the replication stage, 168 SNPs with p value less than 1×10^{-4} in the first stage based on $Z_{1/2}$ were tested using 804 patients and 1470 control samples. We identified additional 234 SNPs using MIN2 in the first stage which could be studied in the replication stage if MIN2 was used instead of $Z_{1/2}$ since MIN2 produces stronger evidence for the additional SNPs than $Z_{1/2}$ does. The Manhattan plots of using MIN2 and $Z_{1/2}$ are presented in Figure 3. One example is *rs*385272 located in chromosome 2, which had a p value of 1.37×10^{-7} which reached significance level at Bonferroni correction in discovery samples alone, whereas $Z_{1/2}$ yielded a p value greater than 1×10^{-4}. Even though there is possibility of false positive findings, these SNPs could have been selected for replication if robust methods were used.

Since we do not have replication data for these additional SNPs selected using MIN2 because the first stage selection was based on $Z_{1/2}$ in Yoon et al. [25], just for illustration purpose of the proposed methods, we present the results of three SNPs including *rs*2131877 that was reported by Yoon et al. [25]. When the significance level in the discovery stage is set at $\alpha_D = 5 \times 10^{-5}$ so that all these exemplary SNPs can be selected in the discovery stage; the p value of combined test based on four robust methods (MAX3, MIN2, GMS, and GME) as well as $Z_{1/2}$ and Pearson's Chi-square test is presented in Table 2. Fisher's combination was used for the joint test in the second stage. Only *rs*2131877 was found to be significant with Bonferroni correction (p value $< 2.03 \times 10^{-7}$) by all except MAX3 method.

5. Discussion

In genetic association studies, efficiency robust tests whose performance does not depend on the underlying genetic model have been extensively studied, and their power benefit over a wide range of genetic models has been well recognized. In this paper, we described how the idea of these robust association tests can be applied to the replication studies and further how overall statistical significance can be evaluated using the combined test formed by p values of the discovery and replication studies.

When the robust tests are used, the test statistic of each stage can have a complex form and thus dealing with the distribution of the joint test can be difficult, whereas calculating the p value of each stage might be relatively simple. Because the asymptotic distribution of the p value under the null hypothesis of no association is easy to handle, the combined test using p values rather than the test statistics themselves can provide computational convenience.

There are several methods for combining test statistics from two stages and Won et al. [22] compared the performances of various choices. Two most commonly used forms are based on Fisher's combination and the linear combination after the inverse normal transformation [23], and we presented the test statistics and p values of these two methods. In our limited experience, the linear combination and Fisher's combination are fairly comparable. Fisher's combination seems to perform slightly better than the linear combination when there exists some heterogeneity between stages in terms of the genotype relative risk, while the linear combination seems to perform slightly better in most of other situations. However, the difference is extremely minor. Further research is required for the thorough comparison of various methods of combining p values in the application of efficiency robust tests to the replication of genetic association studies.

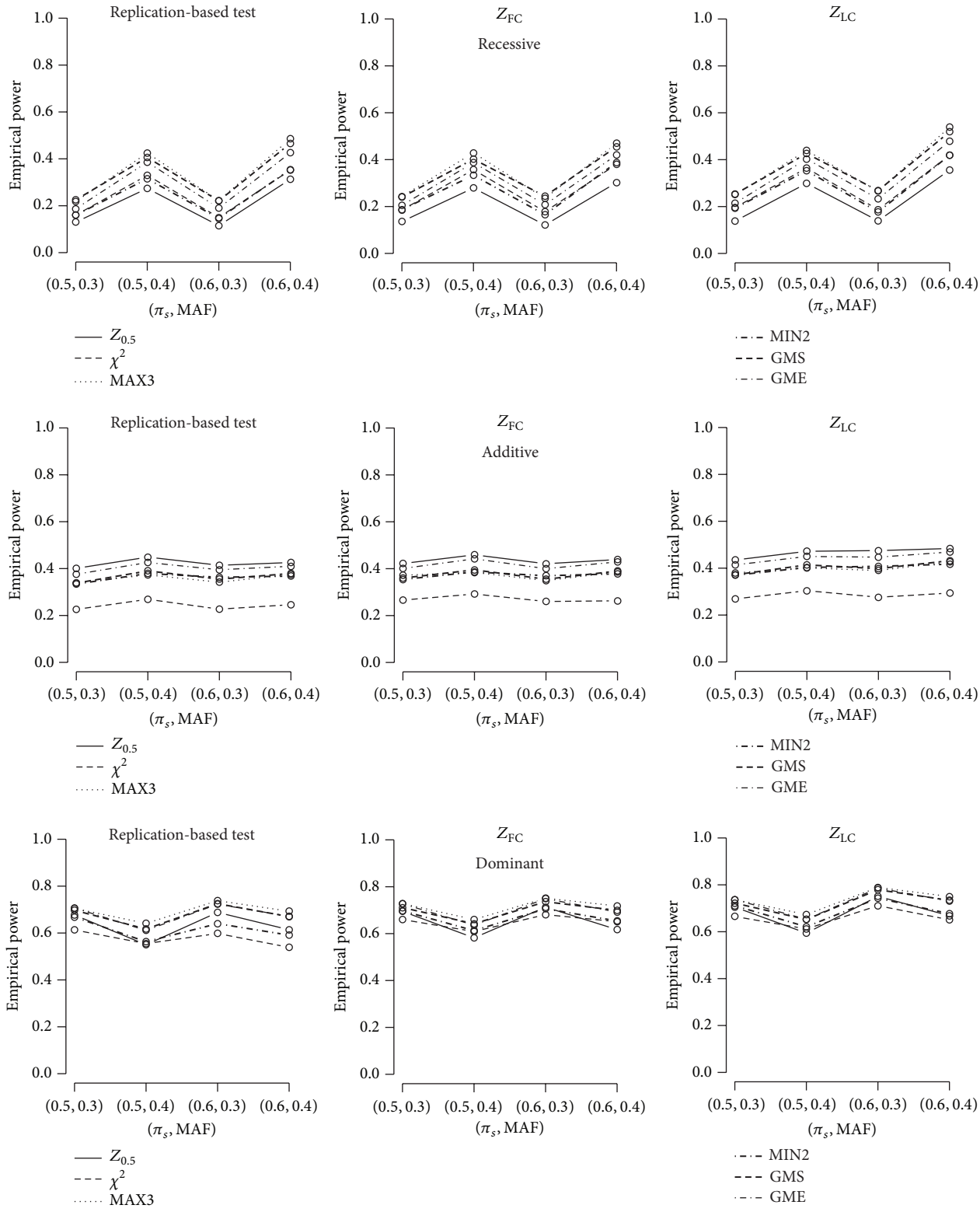


Figure 1: Empirical powers based on 2,000 simulations for $M = 10$ markers, genotype relative risks of both stages = 1.4, and disease prevalence $K = 0.1$ under the recessive, additive, and dominant models. 1,000 cases and 1,000 controls are considered to control $\alpha = 0.05$. The first stage type I error rate for discovery is $\alpha_D = 0.05$. Six test statistics, $Z_{1/2}$, χ^2, MAX3, MIN2, GMS, and GME, are considered. The first, second, and third columns depict powers using the replication-based test, Z_{FC}, and Z_{LC}, respectively.

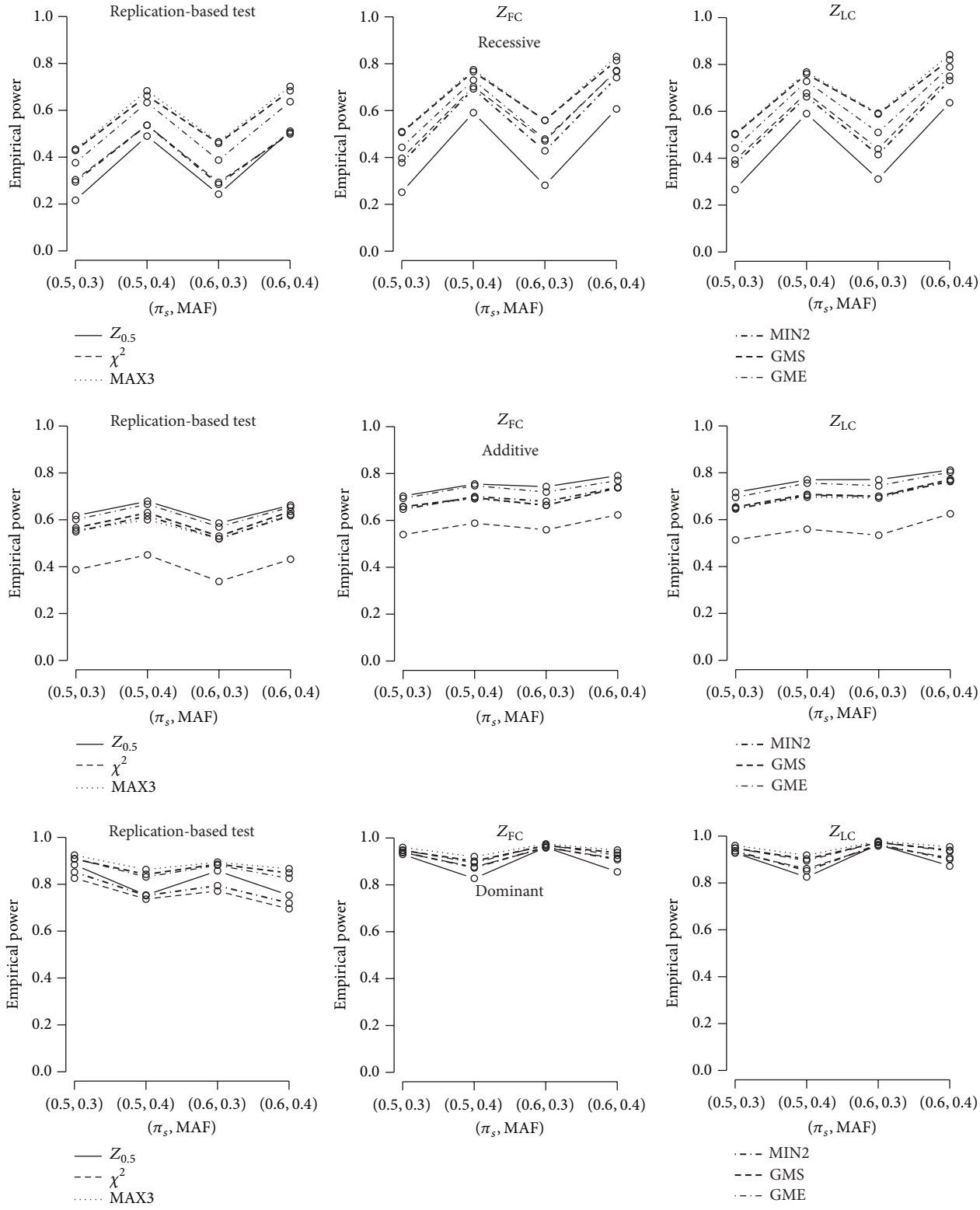


FIGURE 2: Empirical powers based on 2,000 simulations for $M = 10$ markers; genotype relative risks of two stages are different ($r_1 = 1.6$, $r_2 = 1.4$); disease prevalence $K = 0.1$ under the recessive, additive, and dominant models. 1,000 cases and 1,000 controls are considered to control $\alpha = 0.05$. The first stage type I error rate for discovery is $\alpha_D = 0.05$. Six test statistics, $Z_{1/2}$, χ^2, MAX3, MIN2, GMS, and GME, are considered. The first, second, and third columns depict powers using the replication-based test, Z_{FC}, and Z_{LC}, respectively.

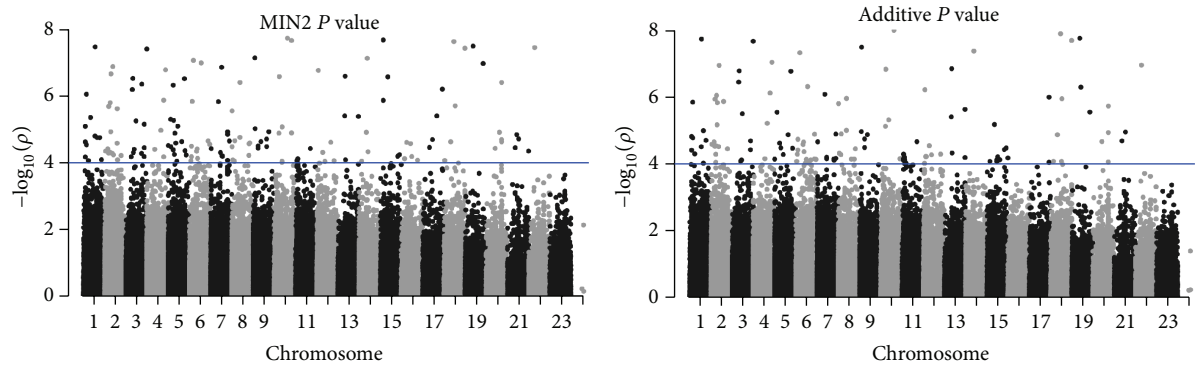


Figure 3: Manhattan plots of 246,758 SNPs from Yoon et al. [25] based on MIN2 (a) and $Z_{1/2}$ (b). The x axis is chromosomal location and the y axis is the significance ($-\log_{10}P$) of association. The horizontal line corresponds to the significance level 10^{-4}.

Table 2: For selected exemplary three SNPs for testing association with NSCLC, p value of combined test using additive CATT ($Z_{1/2}$), Pearson's Chi-square test (T_{chi2}), MAX3, MIN2, Z_{GMS}, and Z_{GME}.

SNP	p value of $Z_{1/2}$			p value of T_{chi2}		
	Discovery	Replication	Combined test	Discovery	Replication	Combined test
rs2131877	7.88×10^{-5}	1.04×10^{-4}	7.97×10^{-8}	1.40×10^{-4}	1.49×10^{-4}	1.84×10^{-7}
rs905551	1.83×10^{-5}	7.02×10^{-3}	7.70×10^{-6}	8.06×10^{-5}	4.89×10^{-2}	1.40×10^{-5}
rs1695109	2.48×10^{-4}	3.46×10^{-2}	2.17×10^{-6}	4.56×10^{-5}	1.53×10^{-1}	2.07×10^{-5}
SNP	p value of MAX3			p value of MIN2		
	Discovery	Replication	Combined test	Discovery	Replication	Combined test
rs2131877	1.53×10^{-4}	4.05×10^{-2}	1.92×10^{-5}	1.32×10^{-4}	1.04×10^{-4}	1.26×10^{-7}
rs905551	4.50×10^{-5}	7.02×10^{-3}	1.92×10^{-6}	1.34×10^{-4}	4.89×10^{-2}	1.99×10^{-5}
rs1695109	3.54×10^{-5}	2.63×10^{-2}	4.64×10^{-6}	2.36×10^{-5}	2.63×10^{-2}	3.35×10^{-6}
SNP	p value of Z_{GMS}			p value of Z_{GME}		
	Discovery	Replication	Combined test	Discovery	Replication	Combined test
rs2131877	1.86×10^{-4}	1.04×10^{-4}	1.71×10^{-7}	1.03×10^{-4}	1.04×10^{-4}	1.02×10^{-7}
rs905551	5.19×10^{-5}	7.02×10^{-3}	2.16×10^{-6}	7.35×10^{-5}	8.01×10^{-3}	3.20×10^{-6}
rs1695109	6.89×10^{-4}	1.27×10^{-1}	3.85×10^{-5}	2.69×10^{-5}	4.19×10^{-2}	5.40×10^{-6}

In a genetic study where the purpose of considering a replication stage is to validate or replicate the genetic findings from the discovery stage, which is the case considered in this paper, the analysis in the replication stage utilized the test statistic or genetic model that is selected as being the best in the discovery stage and also the direction of the risk allele, following guidelines for exact replication in genetic association studies. If the purpose is to simply combine the evidence from different data sources such as in meta-analysis, other strategies may be devised. Further research, again, is required to provide fully detailed properties of such methods.

Power gain of a joint analysis over the conventional replication-based analysis was thoroughly studied by Skol et al. [18, 19]. In our simulation, the amount of power increase using a joint test compared to the replication-based analysis was much minor than what was observed by Skol et al. [18, 19]. The exact reason is not known, but we suspect this might be due to the power advantages of robust methods and also due to the fact that the optimal choice from the first stage is used when calculating the second stage p values. However, even though it was minor in some situations, the joint anlysis presented better power performance than the replication-based analysis in our study. This type of joint analysis raised concerns about the exact meaning of replication [17]. However, McCarthy et al. [26] mentioned that joint analyses "blur the boundaries of where exactly replication starts, but whichever analytical approach is taken, confirmation in many independent samples is important and it is the overall strength of the evidence of association that matters." Purpose of the current study was to present how the overall strength of the evidence of association can be evaluated when robust tests are used in GWAS replication studies.

We illustrated how the proposed methods can be applied in the real data that studied the association of SNPs with non-small-cell lung cancer (NSCLC) in discovery and replication stages. In the original study reported by Yoon et al. [25], SNPs were selected in the discovery data set not based on the robust tests but based on additive CATT. Therefore, we found that some SNPs could have been selected by one

of the robust methods but they were not included in the replication data set. For these SNPs, we were not able to perform the joint analysis that we propose, and it was not possible to examine whether there are other SNPs that could have been found to be associated with NSCLC by proposed methods in the replication study. For this reason, we merely presented how many additional SNPs could have been further followed in the replication stage when robust methods were used. In many GWASs, it is a common practice to report the summary test statistics and p values of the SNPs under a specific genetic model, usually an additive model, which were further genotyped in the replication stage and were finally defined to be significantly associated with a phenotype of interest. As emphasized in this paper, one may have a better chance of finding many missing SNPs by applying more powerful and robust methods that consider different genetic models simultaneously. Therefore, we urge the community to share test results under not only an additive model but also other genetic models, although they were not significant at a stringent significance level, so that future research may have enriched data resources, to which robust tests can be applied in association studies.

Appendix

p value of Fisher's Combination for Equal and Unequal Weights

Equal Weights. Assume $w_1 = w_2 = 1$. Under the null hypothesis of no association, $X_1 = -2\log P^{(1)}$ and $X_2 = -2\log P^{(2)}$ are independent and each asymptotically follows a χ^2 distribution with 2 degrees of freedom Let $f_k(x)$ and $F_k(x)$ be the probability and cumulative density functions of χ^2 random variable with k degrees of freedom Then $f_2(x) = \exp(-x/2)/2$, $f_4(x) = x\exp(-x/2)/4$, $F_2(x) = 1-\exp(-x/2)$, and $F_4(x) = 1-\exp(-x/2)-x\exp(-x/2)/2$. Denote the cutoff of the discovery stage based on α_D as C_D; that is, $F_2(C_D) = 1-\alpha_D$. For observed value $z_{\mathrm{FC}} > C_D$ of $X_1 + X_2$, the p value is written as

$$\begin{aligned} &P_{H_0}\left(X_1 > C_D, X_1 + X_2 > z_{\mathrm{FC}}\right) \\ &= \alpha_D - \int_{C_D}^{z_{\mathrm{FC}}} f_2(x)\, F_2\left(z_{\mathrm{FC}} - x\right) dx \\ &= \alpha_D + \exp\left(-\frac{z_{\mathrm{FC}}}{2}\right) - \alpha_D + \frac{1}{2}\exp\left(-\frac{z_{\mathrm{FC}}}{2}\right)\left(z_{\mathrm{FC}} - C_D\right) \\ &= \exp\left(-\frac{z_{\mathrm{FC}}}{2}\right)\left(1 + \frac{z_{\mathrm{FC}}}{2} + \log\alpha_D\right). \end{aligned} \tag{A.1}$$

Unequal Weights. When different proportions of samples are used in the discovery and replication stages, it may be more appropriate to assign weights proportional to the sample sizes for each stage. For example, when only a small portion is used in the discovery stage, to prevent Fisher's combination test from being dominated by the significant result in the discovery stage, one may want to assign a small weight to the discovery stage result.

When π_s is the proportion of samples used in the discovery stage, one selection for weights is $w_1 = 2\pi_s$ and $w_2 = 2(1-\pi_s)$ for discovery and replication stages, which simplifies to equal weights when $\pi_s = 0.5$. Based on these weights, we consider unequal-weighted Fisher's combination as $-2\log P^{(1)w_1}P^{(2)w_2} = w_1X_1 + w_2X_2$ [27]. Its density function is given by

$$\begin{aligned} f_w(x) = &\frac{1}{2\left(w_1 - w_2\right)}\exp\left(-\frac{x}{\left(2w_1\right)}\right) \\ &- \frac{1}{2\left(w_1 - w_2\right)}\exp\left(-\frac{x}{\left(2w_2\right)}\right), \end{aligned} \tag{A.2}$$

and the probability distribution function is

$$\begin{aligned} F_w(x) = 1 - &\left\{\frac{w_1}{\left(w_1 - w_2\right)}\exp\left(-\frac{x}{2w_1}\right)\right. \\ &\left. - \frac{w_2}{\left(w_1 - w_2\right)}\exp\left(-\frac{x}{2w_2}\right)\right\}, \quad w_1 \neq w_2. \end{aligned} \tag{A.3}$$

Using the previous notation, we have the following form of p value:

$$\begin{aligned} &P_{H_0}\left(X_1 > C_D, w_1X_1 + w_2X_2 > z_{\mathrm{FC}}\right) \\ &= \alpha_D - \int_{C_D}^{z_{\mathrm{FC}}/w_1} f_2(x)\, F_2\left(\frac{z_{\mathrm{FC}} - w_1x}{w_2}\right) dx \\ &= \exp\left(-\frac{z_{\mathrm{FC}}}{2w_1}\right) + \frac{w_2}{w_1 - w_2}\exp\left(-\frac{z_{\mathrm{FC}}}{2w_2}\right) \\ &\quad \times \left\{\exp\left(\frac{w_1 - w_2}{2w_1w_2}z_{\mathrm{FC}}\right) - \exp\left(\frac{w_1 - w_2}{2w_1w_2}w_1C_D\right)\right\} \\ &= \frac{w_1}{w_1 - w_2}\exp\left(-\frac{z_{\mathrm{FC}}}{2w_1}\right) \\ &\quad - \frac{w_2}{w_1 - w_2}\exp\left(-\frac{z_{\mathrm{FC}}}{2w_2}\right)\alpha_D^{-(w_1 - w_2)/w_2} \\ &= \frac{w_1}{w_1 - w_2}\exp\left(-\frac{z_{\mathrm{FC}}}{2w_1}\right) \\ &\quad - \frac{w_2}{w_1 - w_2}\exp\left(-\frac{z_{\mathrm{FC}}}{2w_2}\right)\alpha_D^{-(w_1 - w_2)/w_2}, \end{aligned} \tag{A.4}$$

where $z_{\mathrm{FC}}/w_1 > C_D$.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are indebted to late Dr. Gang Zheng for his inspiration and support on their work. This work was supported by grant of the National Cancer Center (no. NCC-1210060).

References

[1] R. J. Klein, C. Zeiss, E. Y. Chew et al., "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[2] R. H. Duerr, K. D. Taylor, S. R. Brant et al., "A genome-wide association study identifies *IL 23R* as an inflammatory bowel disease gene," *Science*, vol. 314, no. 5804, pp. 1461–1463, 2006.

[3] A. Herbert, N. P. Gerry, M. B. McQueen et al., "A common genetic variant is associated with adult and childhood obesity," *Science*, vol. 312, no. 5771, pp. 279–283, 2006.

[4] R. Sladek, G. Rocheleau, J. Rung et al., "A genome-wide association study identifies novel risk loci for type 2 diabetes," *Nature*, vol. 445, no. 7130, pp. 881–885, 2007.

[5] P. R. Burton, D. G. Clayton, L. R. Cardon et al., "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[6] S. H. Kwak, S. H. Kim, Y. M. Cho et al., "A genome-wide association study of gestational diabetes mellitus in Korean women," *Diabetes*, vol. 61, no. 2, pp. 531–541, 2012.

[7] W. G. Cochran, "Some methods for strengthening the common χ^2 tests," *Biometrics*, vol. 10, no. 4, pp. 417–451, 1954.

[8] P. Armitage, "Tests for linear trends in proportions and frequencies," *Biometrics*, vol. 11, no. 3, pp. 375–386, 1955.

[9] P. D. Sasieni, "From genotypes to genes: doubling the sample size," *Biometrics*, vol. 53, no. 4, pp. 1253–1261, 1997.

[10] B. Freidlin, G. Zheng, Z. Li, and J. L. Gastwirth, "Trend tests for case-control studies of genetic markers: power, sample size and robustness," *Human Heredity*, vol. 53, no. 3, pp. 146–152, 2002.

[11] J. L. Gastwirth, "On robust procedures," *Journal of the American Statistical Association*, vol. 61, no. 316, pp. 929–948, 1966.

[12] J. L. Gastwirth, "The use of maximin efficiency robust tests in combining contingency tables and survival analysis," *Journal of the American Statistical Association*, vol. 80, no. 390, pp. 380–384, 1985.

[13] G. Zheng and H. K. T. Ng, "Genetic model selection in two-phase analysis for case-control association studies," *Biostatistics*, vol. 9, no. 3, pp. 391–399, 2008.

[14] J. Joo, M. Kwak, and G. Zheng, "Improving power for testing genetic association in case-control studies by reducing the alternative space," *Biometrics*, vol. 66, no. 1, pp. 266–276, 2010.

[15] J. Joo, M. Kwak, K. Ahn, and G. Zheng, "A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium," *Biometrics*, vol. 65, no. 4, pp. 1115–1122, 2009.

[16] P. Kraft, E. Zeggini, and J. P. A. Ioannidis, "Replication in genome-wide association studies," *Statistical Science*, vol. 24, no. 4, pp. 561–573, 2009.

[17] D. C. Thomas, G. Casey, D. V. Conti, R. W. Haile, J. P. Lewinger, and D. O. Stram, "Methodological issues in multistage genome-wide association studies," *Statistical Science*, vol. 24, no. 4, pp. 414–429, 2009.

[18] A. D. Skol, L. J. Scott, G. R. Abecasis, and M. Boehnke, "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies," *Nature Genetics*, vol. 38, no. 2, pp. 209–213, 2006.

[19] A. D. Skol, L. J. Scott, G. R. Abecasis, and M. Boehnke, "Optimal designs for two-stage genome-wide association studies," *Genetic Epidemiology*, vol. 31, no. 7, pp. 776–788, 2007.

[20] G. Zheng, B. Freidlin, Z. Li, and J. L. Gastwirth, "Choice of scores in trend tests for case-control studies of candidate-gene associations," *Biometrical Journal*, vol. 45, no. 3, pp. 335–348, 2003.

[21] K. Song and R. C. Elston, "A powerful method of combining measures of association and Hardy-Weinberg Disequilibrium for fine-mapping in case-control studies," *Statistics in Medicine*, vol. 25, no. 1, pp. 105–126, 2006.

[22] S. Won, N. Morris, Q. Lu, and R. C. Elston, "Choosing an optimal method to combine *P*-values," *Statistics in Medicine*, vol. 28, no. 11, pp. 1537–1553, 2009.

[23] F. Begum, D. Ghosh, G. C. Tseng, and E. Feingold, "Comprehensive literature review and statistical considerations for GWAS meta-analysis," *Nucleic Acids Research*, vol. 40, no. 9, pp. 3777–3784, 2012.

[24] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC Press, 1993.

[25] K. A. Yoon, J. H. Park, J. Han et al., "A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population," *Human Molecular Genetics*, vol. 19, no. 24, pp. 4948–4954, 2010.

[26] M. I. McCarthy, G. R. Abecasis, L. R. Cardon et al., "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356–369, 2008.

[27] I. J. Good, "On the weighted combination of significance tests," *Journal of the Royal Statistical Society, Series B*, vol. 17, pp. 264–265, 1955.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 852341, 11 pages
http://dx.doi.org/10.1155/2015/852341

Research Article

Clique-Based Clustering of Correlated SNPs in a Gene Can Improve Performance of Gene-Based Multi-Bin Linear Combination Test

Yun Joo Yoo,[1,2] Sun Ah Kim,[1] and Shelley B. Bull[3,4]

[1] *Department of Mathematics Education, Seoul National University, Seoul 151-742, Republic of Korea*
[2] *Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Republic of Korea*
[3] *Prosserman Centre for Health Research, The Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada M5T 3L9*
[4] *Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada M5T 3M7*

Correspondence should be addressed to Yun Joo Yoo; yyoo@snu.ac.kr

Received 14 November 2014; Revised 3 February 2015; Accepted 14 February 2015

Academic Editor: Taesung Park

Gene-based analysis of multiple single nucleotide polymorphisms (SNPs) in a gene region is an alternative to single SNP analysis. The multi-bin linear combination test (MLC) proposed in previous studies utilizes the correlation among SNPs within a gene to construct a gene-based global test. SNPs are partitioned into clusters of highly correlated SNPs, and the MLC test statistic quadratically combines linear combination statistics constructed for each cluster. The test has degrees of freedom equal to the number of clusters and can be more powerful than a fully quadratic or fully linear test statistic. In this study, we develop a new SNP clustering algorithm designed to find cliques, which are complete subnetworks of SNPs with all pairwise correlations above a threshold. We evaluate the performance of the MLC test using the clique-based CLQ algorithm versus using the tag-SNP-based LDSelect algorithm. In our numerical power calculations we observed that the two clustering algorithms produce identical clusters about 40~60% of the time, yielding similar power on average. However, because the CLQ algorithm tends to produce smaller clusters with stronger positive correlation, the MLC test is less likely to be affected by the occurrence of opposing signs in the individual SNP effect coefficients.

1. Introduction

Current genetic association studies aim to identify genetic variants responsible for a disease by investigating associations between single nucleotide polymorphisms (SNPs) and a trait of interest. In a single-SNP approach, each SNP is analyzed individually for the marginal association with the trait. In a multi-SNP approach, a group of SNPs is analyzed together for polygenic model analysis or gene-based analysis to obtain a global statistic for the combined effect of a set of SNPs [1–8]. When the gene is the unit of interest in the association analysis, gene-based analyses can be performed with multimarker methods using multi-SNP genotypes or haplotypes [6, 8–10]. These methods have the potential benefits of reducing genome-wide type I error burden and boosting the power of the study [9].

Most popular multimarker methods have been developed for the analysis of genotypes. In some methods, the marginal effects of individual SNPs are combined to form a global statistic [5, 11]. In others, SNP genotypes are analyzed in a multiple regression model and global statistics are constructed to represent the joint effects of multiple SNPs in a gene [3, 8, 12, 13]. Some multimarker tests such as C-alpha [14], SKAT [11], and CMC tests [15] specifically target rare variants with minor allele frequency (MAF) less than 1%. On the other hand, multimarker tests such as SKAT-C [16] and the test by Curtis [10] can be applied to a combined set of rare (MAF $< 1\%$), low frequency ($1\% \leq$ MAF $< 5\%$), and common (MAF $\geq 5\%$) variants.

Multimarker methods can be roughly divided into two types: linear and quadratic tests [17]. Linear tests are constructed by combining the individual SNP effects in a linear

combination with certain weights [2, 4, 18]. Linear tests can be powerful if the individual SNP effects have the same direction but can lose substantial power if this condition is not met [2, 5, 8, 17]. Since the direction of an effect can be reversed by recoding the genotype, some researchers developed methods to recode the risk and base alleles considering magnitude and direction of pairwise correlations between SNPs [2, 5]. Quadratic tests are constructed as a quadratic form of an effect vector with corresponding weight matrix [5, 11, 16]. Quadratic tests are more robust to the occurrence of effects in opposing directions, but the degrees of freedom can be high if many neutral SNPs are included in the analysis [13].

Yoo et al. [8, 13] proposed the multi-bin linear combination test (MLC), which is a hybrid of linear and quadratic tests, and evaluated its performance for common SNPs [13] and for combinations of common and low frequency SNPs [8]. For the MLC test, SNPs are partitioned into bins or clusters of highly correlated SNPs according to the pairwise linkage disequilibrium (LD) measure r. Then percluster linear combinations of individual SNP effects are combined in a quadratic form [8, 13]. Because of the quadratic component, the MLC test is more robust than linear tests and can have better power than a quadratic test such as the generalized Wald test under realistic causal model scenarios [8, 13].

For SNP clustering, Yoo et al. [8, 13] previously applied an algorithm incorporated in the LDSelect program [19]. LDSelect is designed to select tag SNPs and the cluster partitioning of a gene proceeds by identifying SNPs that capture the effects of other SNPs through LD. Because its greedy algorithm begins with a SNP in LD with the largest number of other SNPs, it tends to first construct one large cluster. However, for the MLC test, clusters with fewer SNPs are less likely to include causal effects in opposing directions and may therefore be more robust. Yoo et al. also showed that the power of the MLC test is better when correlations between SNPs within a cluster are large and positive [13].

In this study, we develop a new clustering algorithm called CLQ that identifies cliques in the network of SNPs. By definition, all pairwise correlations between SNPs in a clique are above a prespecified threshold value. We also incorporate the coding correction algorithm of Wang and Elston [2, 5] into CLQ so that, after recoding, the cliques consist only of positively correlated SNPs. We compare the performance of MLC tests using the previous clustering algorithm, LDSelect, with that using the new CLQ algorithm in terms of power and robustness. For power calculations, we use genotype data from the HapMap Asian population to provide 1000 different realistic LD structures. For one causal and two causal SNP scenarios, we consider all possible causal SNP choices within each gene. Through extensive numerical power calculations for the MLC test under various causal-gene SNP-trait model scenarios, we show that the CLQ algorithm is highly suitable for incorporation into the MLC test.

2. Methods and Materials

2.1. Multi-Bin Linear Combination Test. Suppose m SNPs in a gene are jointly analyzed in a multiple regression. We denote the genotypes of m SNPs as $X_1, X_2, \ldots, X_m$. The genotypes can be coded differently depending on the genetic model. For the rest of the paper, we assume an additive genetic model such that X_i is the count of minor alleles for ith SNP; that is, $X_i = 0$, 1, or 2. We set up the regression model as

$$g^{-1}\{E(Y)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m, \quad (1)$$

where $g^{-1}(\cdot)$ is the link function. For the global hypothesis of gene-based association, we construct a test using the estimated beta coefficients, $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m)$, and the covariance matrix of $\widehat{\beta}$, Σ.

Suppose m SNPs are partitioned into several bins or clusters based on the pairwise linkage disequilibrium measure r defined as

$$r_{ij} = \frac{P_{ij} - P_i P_j}{\sqrt{P_i P_j (1 - P_i)(1 - P_j)}}, \quad (2)$$

where P_i and P_j are the MAF values of the ith and jth SNP and P_{ij} is the frequency of the haplotype consisting of the two minor alleles. If phase information of genotypes to identify haplotypes is not available, P_{ij} is estimated using maximum likelihood methods, which is the same as computing the Pearson correlation coefficient r between additive genotypes X_i and X_j. If m SNPs are partitioned into l clusters, we construct a $m \times l$ matrix J to denote SNP assignments such that $J_{ij} = 1$ if the ith SNP belongs to the jth cluster and $J_{ij} = 0$ if not.

Using the assignment matrix J, we construct an $l\ df$ MLC test statistic such that

$$G_M = \left(W^T \widehat{\beta}\right)\left(W^T \Sigma W\right)^{-1}\left(\widehat{\beta}^T W\right), \quad (3)$$

where $W = (\Sigma^{-1}J)(J^T\Sigma^{-1}J)^{-1}$ [8, 13].

If only one SNP is assigned to each cluster (singleton), G_M is the same as the generalized Wald test statistic

$$G_W = \widehat{\beta}^T \Sigma^{-1} \widehat{\beta}. \quad (4)$$

Moreover, if all SNPs are assigned to one cluster, G_M is a linear combination (LC) test [1]. The asymptotic null distribution of the Wald test statistic is an $m\ df$ chi-square distribution, assuming no linear dependencies among SNP genotypes, whereas the MLC test statistic follows an $l\ df$ chi-square distribution. The asymptotic null distribution of the LC test statistic is 1 df chi-square.

2.2. Allele Recoding Algorithm. As shown in Yoo et al. [13], power of the LC and the MLC tests benefits from high positive correlation between causal and neutral SNPs or between causal SNPs. The sign of the correlation r_{ij} between two SNPs changes if we switch the risk and base alleles for one of the two SNPs. For example, if we replace X_i with a new genotype variable $X_i' = 2 - X_i$, then the genotype of $X_i = 0, 1, 2$ becomes $X_i' = 2, 1, 0$, respectively, under an additive model. When this change is applied, the correlation between the genotype X_i' and X_j becomes $-r_{ij}$ for $i \neq j$. This coding change will also

change the sign of beta estimates $\widehat{\beta}_i$ if $\widehat{\beta}_i \neq 0$. To make most pairwise correlations positive for SNPs in the joint analysis, we apply the Wang and Elston's SNP recoding algorithm [2], which is as follows.

Step 1. Count the number of negatively correlated SNPs for each SNP i and denote it as n_i for $i = 1, 2, \ldots, m$; that is, $n_i = \sum_{j=1, j\neq i}^{m} I\ (r_{ij} < 0)$, where I is an indicator function.

Step 2. Select the SNP with the $\max\{n_i\}$; then switch the risk and base allele for the genotype of that SNP.

Step 3. Iterate Steps 1-2 with updated correlations from the updated genotypes until $\max\{n_i\} < m/2$.

For the MLC test based on the LDSelect algorithm, we applied recoding for each cluster separately after clustering. With the CLQ algorithm, we incorporate recoding within the clustering algorithm.

2.3. SNP Clustering Using the LDSelect Algorithm. The LDSelect algorithm [19] is as follows.

Step 1. Set a threshold value c for correlation r between two SNPs. Suppose the m SNPs in a gene are indexed with $V_1 = \{1, 2, \ldots, m\}$. Let $V' := V_1$.

Starting with B_1, iterate the selection of the kth cluster B_k in Steps 2 to 4.

Step 2. For each SNP i in V', count the number of other SNPs having correlation with SNP i greater than a threshold value c such that $t_i = \sum_{j\in V', j\neq i} \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \left|r_{ij}\right| > c \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We call the SNPs that meet this criteria the *neighbors* of SNP i. Proceed to Step 5 if t_i is at most equal to 0 for all $i \in V'$. If not, proceed to the next step.

Step 3. First, select one SNP (say j) with $t_j = \max_{i\in V'} t_i$ and all the neighbors of SNP j in V' and group them as $B_k = \{i \in V' : |r_{ij}| > c\}$. When there is more than one SNP with the maximum number of neighbors, randomly select one SNP from among them.

Step 4. Remove SNPs in B_k from V_k and denote it as $V_{k+1} = V_k \setminus B_k$. Also, update V' with V_{k+1}. Iterate Steps 2~4 unless the condition to proceed to Step 5 is met or all SNPs are assigned into a cluster.

Step 5. If the maximum t_i for all $i \in V'$ is at most 0, the SNPs in V' will be partitioned into singleton clusters (each with only one SNP).

End. In this way all the SNPs are assigned into clusters $B_1, \ldots, B_l$ for some $l \leq m$. Then $V_1 = \cup_{k=1}^{l} B_k$, where $B_j \cap B_k = \phi$ for $j \neq k$ and $B_k \neq \phi$ for $k = 1, \ldots, l$.

2.4. SNP Clustering Using CLQ Algorithm. Let $G = (V, E)$ be a graph with a vertex set V and an edge set E of V, the set of some pairs of vertices in V. If an edge between two vertices is included in E, we call these two adjacent. A *clique* is defined as a subset C of V such that all pairs of vertices in C are adjacent. A *maximal clique* in G is a clique whose vertices are not a subset of the vertices of a larger clique, and the *maximum cliques* in G are the largest among all cliques in a graph. A *subgraph* of G is a graph with a vertex set $V' \subseteq V$ and an edge set $E' \subseteq E$. A subgraph $G' = (V', E')$ of G is said to be *induced* by a vertex set $V' \subseteq V$ when an edge is in E' if and only if the edge is in E. The subgraph induced by a clique is complete (all possible edges between vertices in clique are included in the edge set).

The CLQ algorithm is as follows.

Step 1. For a threshold value c, construct a graph G_1 with a vertex set $V_1 = \{1, 2, \ldots, m\}$ corresponding to SNP $1, 2, \ldots, m$ in a gene and an edge set E_1 in which the undirected edge between vertex i and j is included if $|r_{ij}| > c$ for $i \neq j$. Let $G' := G_1$ with $V' := V_1$ and $E' := E_1$.

Starting with B_1, iterate the selection of the kth cluster B_k in Steps 2 to 4.

Step 2. For each vertex in G', find the maximal cliques that contain the vertex using the Bron-Kerbosch algorithm [20] implemented in igraph package [21] and select the largest clique of maximal cliques found for all vertices. Proceed to Step 5 if there is no maximum clique with at least two vertices. Otherwise, proceed to the next step.

Step 3. Apply the recoding algorithm to the maximum cliques chosen in Step 2. If all pairwise correlations between SNPs in the clique can be recoded to be positive, then take the SNPs corresponding the chosen clique as the cluster B_k. If negatively correlated SNPs still exist after the recoding algorithm has been applied to this clique, discard the chosen clique and select the next largest one. If there are multiple cliques in G' with the largest size and all SNPs can be recoded to be positively correlated, choose the one with the largest sum of absolute correlation. Repeat application of the recoding algorithm until B_k is determined. If there is no clique with at least two vertices that can be recoded to have all positive correlations, proceed to Step 5.

Step 4. Remove SNPs in B_k from V_k and denote it as $V_{k+1} = V_k \setminus B_k$. Also, update V' with V_{k+1}. Update G' with the subgraph G_{k+1} induced by V_{k+1}. The edge set is also updated by the edge set of G_{k+1} as $E' := E_{k+1}$. Iterate Steps 2~4 unless the condition to proceed to Step 5 is met or all SNPs are assigned into a cluster.

Step 5. If there is no maximum clique with at least two vertices in G', the SNPs in V' will be partitioned into singleton clusters.

End. In this way all the SNPs are assigned into clusters $B_1, \ldots, B_l$ for some $l \leq m$. Then $V_1 = \cup_{k=1}^{l} B_k$, where $B_j \cap B_k = \phi$ for $j \neq k$ and $B_k \neq \phi$ for $k = 1, \ldots, l$.

2.5. Comparison of Clustering Results. To compare the clusters produced for a gene by the two different clustering methods, we used the S criterion of Rand [22] and the S' criterion that is adjusted for chance agreement [23]. Suppose in one clustering method, m, that SNPs are partitioned into $B_1, \ldots, B_l$ and, using another method, they are partitioned into $C_1, \ldots, C_h$. The similarity between two clustering results is defined as

$$S = \frac{\left(\sum_{i<j}^{m} a_{ij}\right)}{\binom{m}{2}}, \tag{6}$$

where $a_{ij} = 1$ if there exist k and k' such that SNP i and SNP j are both in B_k and $C_{k'}$, or if there exist k and k' such that SNP i is in both B_k and $C_{k'}$ while SNP j is in neither B_k nor $C_{k'}$, and $a_{ij} = 0$ otherwise. A higher value of S corresponds to more similar performance of two clustering methods for the given data. When a pair of clustering results are exactly identical, then $S = 1$, whereas if $S = 0$ there is no similarity. The adjusted agreement measure S' is defined as

$$S' = \frac{\sum_{i,j}\binom{m_{ij}}{2} - \sum_i \binom{m_{i\cdot}}{2}\sum_j \binom{m_{\cdot j}}{2} / \binom{m}{2}}{(1/2)\left[\sum_i \binom{m_{i\cdot}}{2} + \sum_j \binom{m_{\cdot j}}{2}\right] - \sum_i \binom{m_{i\cdot}}{2}\sum_j \binom{m_{\cdot j}}{2} / \binom{m}{2}}, \tag{7}$$

where m_{ij} denotes the number of common SNPs that belong to clusters B_i and C_j, and $m_{i\cdot}$ and $m_{\cdot j}$ are the number of SNPs in clusters B_i and C_j, respectively [23].

2.6. Numerical Power Analysis Based on HapMap Data. Based on 1000 gene structures obtained from HapMap phase III Asian data, we computed MLC test power using LDSelect clustering (MLC-LD) and CLQ clustering (MLC-CL) for several alternative trait models with one or two causal SNPs. The HapMap gene panels were obtained by random selection from the 8883 genes that had 4~30 SNPs, excluding SNPs with MAF < 0.01 or any pairwise LD value $|r| > 0.99$. Two sets of 1000 genes were randomly selected, allowing overlap: one for the analysis of Models A, B, and C and another for validation analysis. For each set of 1000 genes, two panels of SNPs with MAF ≥ 0.05 and MAF ≥ 0.01, respectively, were used in comparisons of clustering results, and one panel with MAF ≥ 0.01 was used for power analysis. We evaluated a range of clustering threshold values for c equal to 0.3 through 0.9 for LDSelect and CLQ.

For trait models, we considered models with one to four causal SNPs within a gene and a linear model for the quantitative phenotype Y:

$$Y = \sum_{i=1}^{t} b_i G_i + \varepsilon, \tag{8}$$

where t is the number of causal SNPs, b_i is the effect of ith causal SNP, G_i is the number of causal alleles for the ith causal SNP, and ε is the error term assumed to follow a normal distribution with mean 0 and variance σ^2 (Table 1).

Initially we considered three types of trait models: one with one causal SNP in a gene with effect $b_1 = 1$ (Model A), another with two causal SNPs in a gene with effects $b_1 = 1$, $b_2 = 1$ (Model B), and a third with two causal SNPs in a gene with effects $b_1 = 1$, $b_2 = -1$ (Model C). Since power in a linear model depends on the ratio of signal to noise, that is, (b_i/σ), we selected the variance σ^2 to correspond to reasonable power for a given gene structure and choice of causal SNPs for Models A, B, and C. To clearly see relative performance of the MLC tests compared to the generalized Wald test, we adjusted σ^2 such that the Wald test power is 60% for each trait model. For Model A, in each gene we chose each of the SNPs in turn to be the causal SNP, resulting in over 11,000 trait model settings in total over 1000 genes. For Models B and C, in each gene we chose each of all possible SNP pairs in turn to be the causal SNP pair, resulting in nearly 80,000 trait model settings in total for main and validation sets.

Table 1: Quantitative trait models used for power comparisons of MLC-LD and MLC-CL.

Model name	Description	Trait model parameters*
Model A	One causal SNP within a gene	$b_1 = 1$
Model B	Two causal SNPs, both deleterious	$b_1 = 1, b_2 = 1$
Model C	Two causal SNPs, one deleterious and one protective	$b_1 = 1, b_2 = -1$
Model D	1~4 causal SNPs, random assignment of the direction of effects	$\lvert b_i\rvert$ is randomly selected from the $U(0.01 \times \mathrm{SD}, 0.05 \times \mathrm{SD})$ where SD is the expected standard deviation of Y

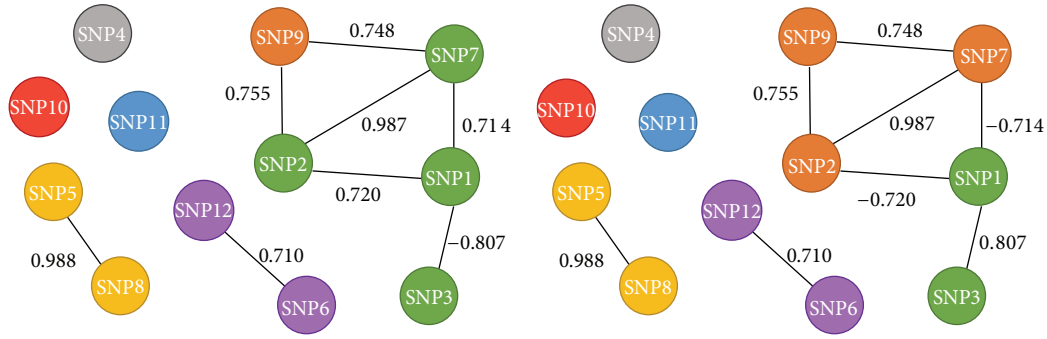
*The trait model is $Y = \sum_{i=1}^{C} b_i G_i + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, C is the number of causal SNPs, b_i is the effect of ith causal SNP, and G_i is the number of causal alleles for the ith causal SNP. The variance σ^2 is adjusted to make the power of Wald test 60% for each set of causal SNPs for Models A, B, and C and set to 1 for Model D.

In a fourth trait model (Model D), we also obtained power over mixed types of causal models with variable effect sizes. The number of causal SNPs was chosen randomly between 1 and 4, with the deleterious and protective causal SNPs randomly assigned. For each causal SNP, $|b_i|$ was randomly selected from the uniform distribution $U(0.01\times\mathrm{SD}, 0.05\times\mathrm{SD})$ where SD is the expected standard deviation of Y following the effect size estimates for SNPs associated with lipid levels presented in Willer et al. [24]. Then the error variance σ^2 was fixed as 1.

For power analysis using Models A to D, the genotype data were randomly generated from the haplotype panel of HapMap data for $n = 5,000$ subjects. Power was calculated numerically for each gene assuming asymptotic chi-square distributions under the null and alternative trait models. With a given significance level α and number of clusters l, the critical value $c_{l,\alpha}$ is obtained from the l df chi-square distribution such that $P\{\chi_l^2 > c_{l,\alpha}\} = \alpha$. Power is computed as $P\{\chi_{l,\lambda}^2 > c_{l,\alpha}\}$ with l df and noncentrality λ parameter equal

Table 2: Mean and standard deviation over 1000 genes of agreement measure S and S' for two clustering methods (LDSelect and CLQ) and number of genes with identical clustering.

Allele frequency cut	c	S		S'		Cases of perfect agreement
		Mean	SD	Mean	SD	
.05	.3	.676	0.203	.325	0.342	180
	.4	.769	0.191	.510	0.336	283
	.5	.847	0.168	.665	0.303	388
	.6	.909	0.123	.781	0.242	483
	.7	.936	0.101	.832	0.210	541
	.8	.959	0.086	.884	0.178	648
	.9	.974	0.069	.918	0.156	736
.01	.3	.689	0.196	.395	0.376	155
	.4	.789	0.177	.559	0.379	254
	.5	.863	0.151	.687	0.338	361
	.6	.923	0.105	.794	0.267	468
	.7	.948	0.084	.843	0.230	536
	.8	.968	0.068	.892	0.201	644
	.9	.981	0.053	.922	0.172	744



(a) LDSelect (b) CLQ

Figure 1: Clustering of gene ARHGAP29 by LDSelect and CLQ for a threshold value 0.7.

to the expected MLC statistic value under the trait model (see Appendix in [13]).

3. Results

3.1. Comparison of SNP Clustering by LDSelect and CLQ. In Figure 1, we illustrate LDSelect and CLQ clustering for 12 SNPs in ARHGAP29 at a threshold value 0.7. By LDSelect, the largest cluster includes SNPs 1, 2, 3, and 7 since SNP 1 tags SNPs 2, 3, and 7. However with CLQ, these four SNPs do not form a clique because the pairwise correlations between SNPs 2 and 3 and between SNPs 3 and 7 are below the threshold value. By CLQ, SNPs 2, 7, and 9 are clustered as a clique and SNPs 1 and 3 are clustered as another clique. Here, SNP 1 is recoded so that the correlation within the clique is positive.

We compared LDSelect and CLQ clustering for each of the 1000 HapMap genes across a range of threshold values. For a given threshold, the clustering methods often produce identical gene clusters, particularly at higher threshold values (Table 2). For example, at the threshold value 0.7, 54% of the clustering results are the same. With increased threshold values, the averages of the agreement measures S and S' also increase. At threshold values greater than 0.5, the average agreement measure S' is greater than 78% overall. Comparison of average S and S' under stratification by five gene-size groups (≤10, 11~15, 16~20, 21~25, >25 SNPs per gene) showed that the agreement slightly weakens with increased number of SNPs (results not shown).

On average, the number of clusters obtained by LDSelect is usually smaller than that by CLQ for a given threshold value (Table 3). Cluster sizes are smaller and less variable in CLQ than in LDSelect, averaged over all gene sizes (Table 3). Figure 2(a) shows the average over 1000 genes of the ratio of the number of clusters to the number of SNPs per gene used for clustering by LDSelect and CLQ given a threshold value c. Restricting the SNPs to have higher minor allele frequency (MAF ≥ 0.05 versus MAF ≥ 0.01) reduces the ratio similarly for both clustering methods. At the same threshold value, CLQ produces a larger number of clusters compared to LDSelect, mainly because CLQ has a stricter within-cluster LD requirement, but this difference decreases as the threshold value increases. It follows that the average

Table 3: The average over 1000 genes of the number of clusters per gene, the mean size of the clusters within a gene, and the standard deviation of the cluster sizes within a gene for two clustering methods (LDSelect and CLQ).

Allele frequency cut	c	# of clusters*		Mean size of clusters*		SD size of clusters*	
		LDSelect	CLQ	LDSelect	CLQ	LDSelect	CLQ
.05	.3	1.84	2.94	6.39	3.70	*2.55*	*2.60*
	.4	2.43	3.27	4.82	3.40	2.82	2.33
	.5	3.02	3.67	3.85	3.06	2.49	2.02
	.6	3.65	4.19	3.13	2.67	2.13	1.75
	.7	4.36	4.79	2.61	2.32	1.75	1.46
	.8	5.18	5.52	2.18	2.01	1.38	1.19
	.9	6.29	6.57	1.76	1.65	1.00	0.89
.01	.3	2.40	3.53	5.25	3.28	3.33	2.53
	.4	3.02	3.91	4.08	3.00	3.02	2.25
	.5	3.66	4.36	3.32	2.71	2.52	1.96
	.6	4.35	4.90	2.75	2.40	2.07	1.67
	.7	5.10	5.57	2.34	2.10	1.66	1.40
	.8	5.99	6.32	1.98	1.85	1.31	1.15
	.9	7.17	7.42	1.64	1.56	0.94	0.85

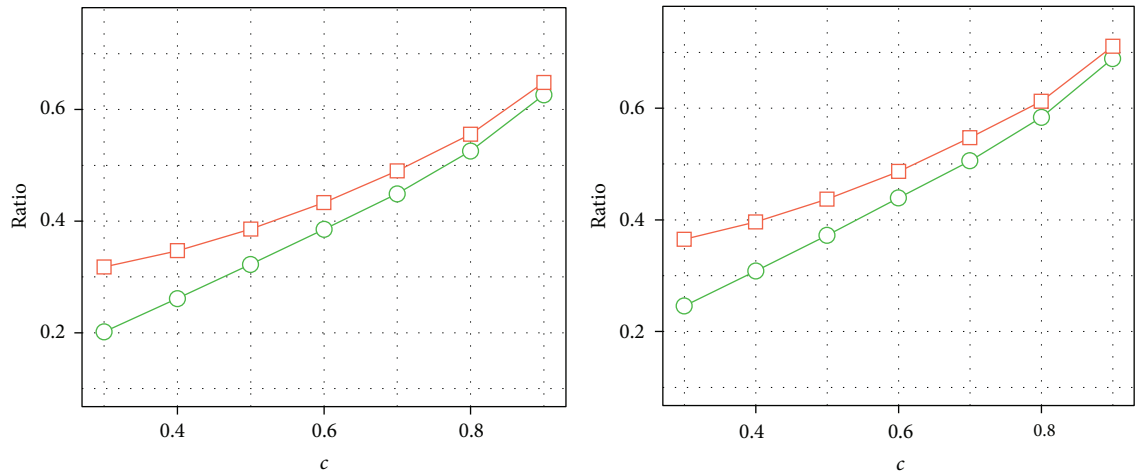
*The differences of the obtained characteristics within genes are compared by paired t-test and all results were significant with P values $<1e^{-10}$ except the italic pairs ($P = 0.61$).

size of the largest cluster in each gene is greater for LDSelect than for CLQ (Figure 2(b)), with greater differences at lower threshold values. Conversely, CLQ usually produces more singleton clusters than LDSelect (Figure 2(c)). We conclude that at the same threshold value, CLQ tends to produce more clusters of smaller size than LDSelect.

To compare maximum cluster size when the number of clusters per gene is the same, rather than at a fixed threshold value, we applied the clustering methods across a range of threshold values and for each gene matched the LDselect and CLQ clustering results according to the number of clusters (Figure 3(a)). At nearly all cluster numbers, the average size of the largest cluster is again smaller for CLQ. Out of all clustering results with different numbers of clusters, 69% had the same maximum cluster size by LDSelect and CLQ, 23% had a larger maximum cluster by LDSelect, and only 8% had a smaller maximum cluster size by LDSelect, based on SNPs with MAF $\geq$ 0.05. For the SNPs with MAF $\geq$ 0.01, these percentages were 65%, 28%, and 7%. For a fixed number of clusters, the number of singleton clusters was slightly smaller for CLQ than LDSelect (Figure 3(b)). Out of all clustering results, 70% had the same number of singleton clusters by LDSelect and CLQ, 21% had a larger number by LDSelect, and only 9% had a smaller number of singleton clusters by LDSelect than CLQ, based on SNPs with MAF $\geq$ 0.05. For the SNPs with MAF $\geq$ 0.01, these percentages were 67%, 25%, and 8%. We conclude that when the two clustering methods produce the same number of clusters for a gene, the CLQ clusters will tend to be less variable in size than the LDSelect clusters. We draw similar conclusions from the analysis of validation set (results not shown).

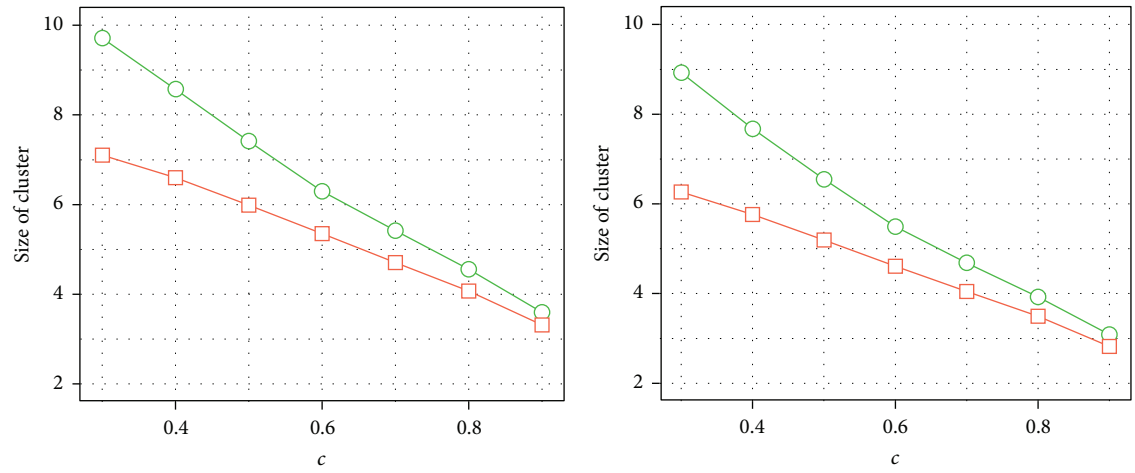
3.2. Comparison of MLC-LD and MLC-CL Test Power. For the power calculations, the variance of the error term was set such that Wald test power is 60% for Models A, B, and C. For Model D, the variance was fixed as 1 and variation in the regression coefficient determined power. In Table 4, the average MLC test power values for trait model types A, B, C, and D using two clustering methods (MLC-LD and MLC-CL) vary across values of the clustering threshold c. When averaged over all sets of genes and causal SNP choices, MLC-LD power and MLC-CL power were both higher than Wald test power (which was 60% for Models A–C and roughly 35–38% for Model D). Average power was usually maximized at $c = 0.6$ or 0.7 for LDSelect and at $c = 0.4$ or 0.5 for CLQ. At threshold values less than 0.7, MLC-CL power was higher than MLC-LD for all models. At threshold values higher than 0.6, however, MLC-CL power was less than MLC-LD for Models A, B, and D. For Model C, MLC-CL power was higher than MLC-LD for all threshold values. For each model, the highest average power was achieved by MLC-CL (bolded entries in Table 4). Comparison of average MLC power values under stratification by five gene-size groups ($\leq$10, 11~15, 16~20, 21~25, >25 SNPs per gene) generally yielded similar results (results not shown).

We also compared the proportion of gene-causal-SNP cases in which MLC-LD power or MLC-CL power was less than Wald test power (Table 4). The proportion with lower MLC-CL power was smaller, suggesting improved robustness. At lower threshold values particularly, the proportion with power less than the Wald test for LDSelect was much higher, up to 40~54% for some models, but was less than 25% for CLQ. Plots of gene-specific power obtained for the cases in which the LDSelect and CLQ clusters differ show that the MLC test using CLQ is less likely than MLC test using LDselect to have substantially reduced power relative to the Wald test (Figure 4). Similar conclusions about relative power were obtained from the analysis of validation set (results not shown).



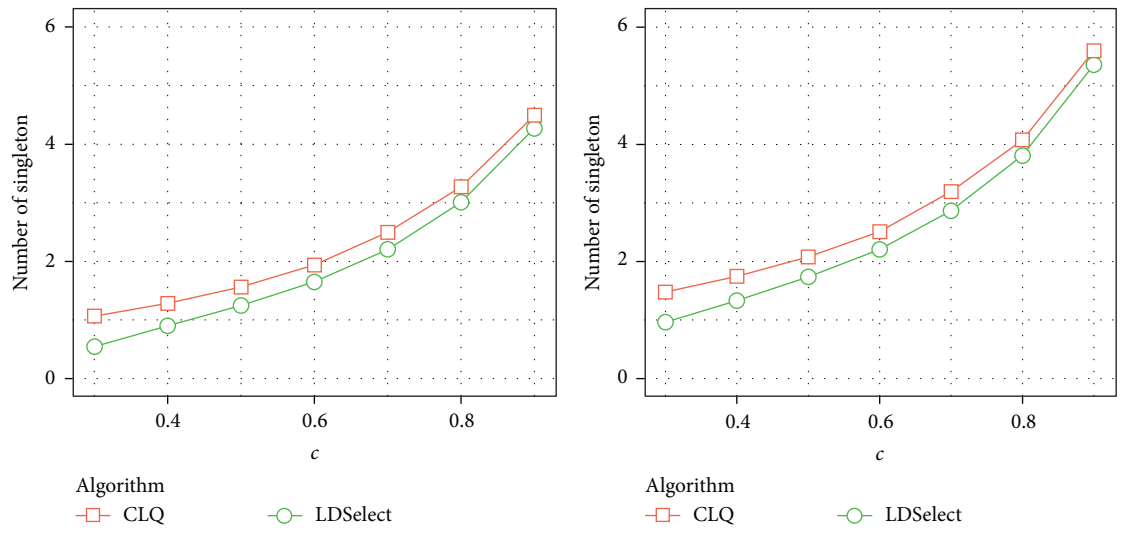
(A) SNPs with allele frequency ≥ 0.05 (B) SNPs with allele frequency ≥ 0.01

(a) The ratio of number of clusters to number of SNPs



(A) SNPs with allele frequency ≥ 0.05 (B) SNPs with allele frequency ≥ 0.01

(b) The size of the largest cluster



(A) SNPs with allele frequency ≥ 0.05 (B) SNPs with allele frequency ≥ 0.01

(c) The number of singleton clusters

Figure 2: Averages of (a) the ratio of number of clusters to number of SNPs, (b) the size of the largest cluster, and (c) the number of singleton clusters in each of 1000 genes for LDSelect and CLQ clustering given a threshold value c.

Table 4: Average MLC test power over all gene-causal-SNP combinations for LDSelect (MLC-LD) and CLQ (MLC-CL) clustering methods and the proportion of genes where MLC-LD power and MLC-CL power are less than Wald test power.

Model	c	All possible causal SNPs and all genes					All possible causal SNPs for the genes where LDSelect and CLQ clusters are different				
		N	Average[†,*]		% Power < Wald[*]		N	Average[†,*]		% Power < Wald[*]	
			LDS	CLQ	LDS	CLQ		LDS	CLQ	LDS	CLQ
A	0.3	11,117	0.627	0.757	36.6	6.2	9,765	0.614	0.759	40.0	5.8
	0.4	11,117	0.670	**0.758**	26.4	3.9	8,867	0.656	**0.762**	30.5	3.3
	0.5	11,117	0.716	0.754	14.6	2.2	8,069	0.714	0.759	17.2	1.8
	0.6	11,117	**0.735**	0.745	6.7	1.0	7,381	0.742	0.753	8.1	0.7
	0.7	11,117	0.733	0.730	2.7	0.6	6,234	**0.751**	0.744	3.4	0.2
	0.8	11,117	0.719	0.712	1.1	0.6	5,138	0.746	0.731	0.8	0.0
	0.9	11,117	0.691	0.685	1.4	1.3	3,512	0.726	0.707	0.3	0.0
B	0.3	79,650	0.645	0.771	33.7	5.6	74,715	0.640	0.774	35.0	5.2
	0.4	79,650	0.682	**0.773**	25.5	3.6	70,384	0.674	**0.775**	27.3	3.0
	0.5	79,650	0.727	0.769	14.5	2.1	66,788	0.723	0.770	15.8	1.7
	0.6	79,650	**0.750**	0.760	6.4	1.2	63,848	0.752	0.764	7.0	0.9
	0.7	79,650	0.748	0.745	3.0	0.6	57,300	**0.756**	0.752	3.5	0.5
	0.8	79,650	0.733	0.724	0.9	0.4	48,577	0.752	0.737	0.7	0.2
	0.9	79,650	0.701	0.692	0.9	0.5	33,403	0.724	0.706	0.8	0.1
C	0.3	79,650	0.499	0.649	54.3	23.7	74,710	0.505	0.663	54.2	21.9
	0.4	79,650	0.551	0.657	44.1	21.1	70,409	0.557	0.675	44.0	18.6
	0.5	79,650	0.603	0.662	32.8	18.4	66,772	0.615	**0.683**	32.0	15.5
	0.6	79,650	0.637	**0.664**	23.7	16.4	63,910	0.651	0.682	22.8	14.1
	0.7	79,650	0.652	0.662	18.1	14.1	57,409	0.669	0.682	17.4	12.2
	0.8	79,650	**0.654**	0.657	14.1	11.7	48,669	**0.675**	0.678	13.8	10.3
	0.9	79,650	0.645	0.646	10.3	8.8	33,625	0.661	0.662	11.6	8.3
D[**]	0.3	8,883	0.388	0.444	36.5	12.1	7,054	0.372	**0.441**	39.7	9.9
	0.4	8,883	0.408	0.447	28.3	9.7	6,140	0.389	0.440	32.5	7.3
	0.5	8,883	0.426	**0.447**	18.8	7.4	5,119	0.404	0.433	22.1	5.0
	0.6	8,883	**0.439**	0.445	10.8	5.5	4,420	**0.425**	0.435	12.5	3.5
	0.7	8,883	*0.439*	*0.440*	6.5	4.2	3,625	*0.416*	*0.414*	7.4	3.0
	0.8	8,883	0.435	0.433	4.4	3.3	2,827	0.419	0.412	4.1	1.7
	0.9	8,883	0.425	0.423	3.6	3.3	2,103	0.406	0.396	2.7	1.7

[*]The differences of power between two clustering algorithm and the proportions of cases with MLC test power less than the power of Wald test within genes are compared by paired t-test and McNemar test, respectively, and all results are significant with P values <0.05 except the italic pairs.

[**]The power of Wald test for Models A, B, and C were fixed as 0.6, whereas the average power of Wald test for Model D was 0.388 in average over all genes (left) and 0.377, 0.373, 0.365, 0.368, 0.354, 0.356, and 0.351 for $c = 0.3 \sim 0.9$, respectively, for genes with clustering results are different (right).

[†]Bolded numbers are the maximum average power of MLC over different threshold values within each clustering method, trait model, and the set of genes (all or the ones with different clustering results by LDSelect and CLQ).

4. Discussion

In previous studies, we reported that power of the MLC test depends on the correlation structure among SNPs and we postulated that clusters of strongly correlated SNPs with positive correlations benefit the test [13]. Therefore, the CLQ algorithm designed to construct such clusters, in which all pairwise correlations in the cluster are positive and strong, should work well for MLC tests. LDSelect and CLQ produced exactly identical clusters for about 38~54% of the genes at threshold values c = 0.5~0.7. This implies that many of the clusters found by LDSelect using a SNP that tags other SNPs are actually cliques; that is, the pairwise correlations between SNPs in the cluster other than the SNP with most neighbors are also above the threshold, even though the LDSelect algorithm does not consider that information.

The LDSelect algorithm was originally developed for tag SNP selection so that indirect associations could be efficiently captured by genotyping and analyzing only tag SNPs. We observed that the LDSelect algorithm also works reasonably well for MLC tests where power depends on formation of the clusters with large positive correlations. Because the algorithms produce identical clusters for a substantial portion of the cases, the average MLC test power values were not

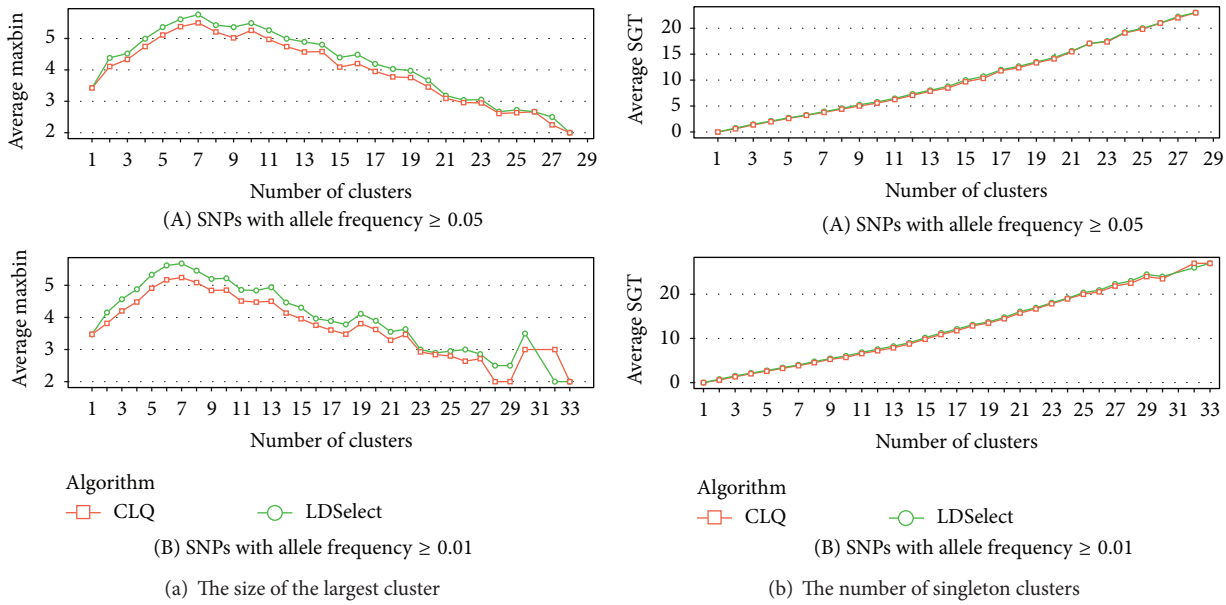


FIGURE 3: Averages of (a) the size of the largest cluster and (b) the number of singleton clusters produced in each gene by LDSelect and CLQ for a fixed number of clusters per gene. For each gene, the number of clusters produced by each clustering method was found at threshold values within a grid from 0.1 to 0.9 by 0.01. Excluding results at the extremes (i.e., including cluster numbers that fell between 20% and 90% of the number of SNPs), the LDSelect and CLQ cluster numbers were matched for each gene and the maximum cluster size for each was averaged across genes at a fixed value of the number of clusters.

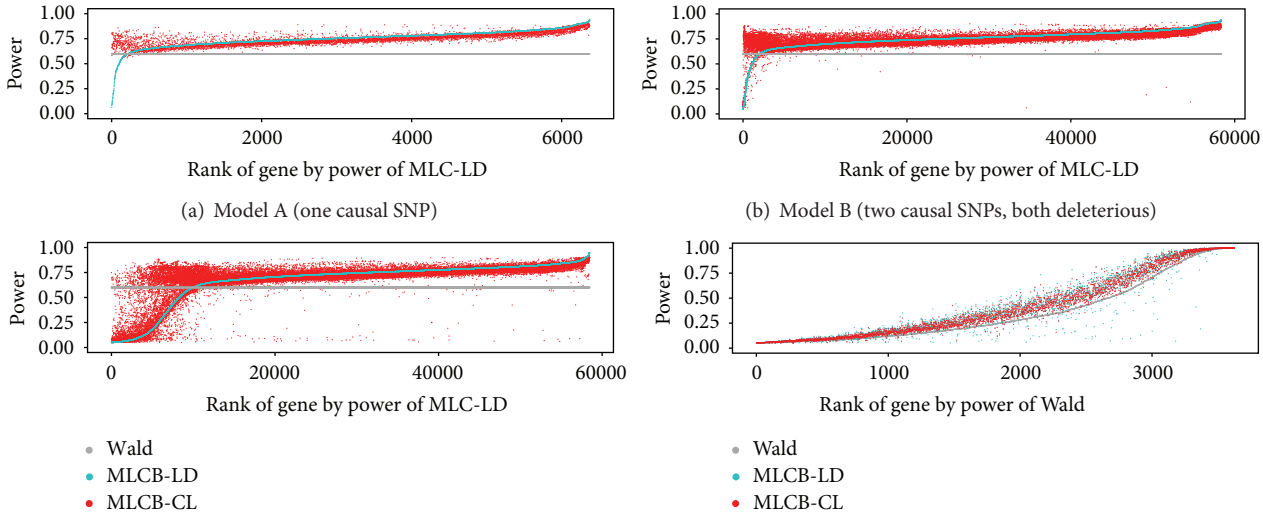


FIGURE 4: Power of MLC test based on LDSelect clustering (MLC-LD: blue) and CLQ clustering (MLC-CL: red) with threshold $c = 0.7$ for the cases where LDSelect and CLQ clusters are different. Each point represents one case of all possible causal SNP assignments within a gene. The variance of the error term was adjusted for each of 1000 genes such that the Wald test power is exactly 0.6 (Grey) for Models A, B, and C and fixed as 1 for Model D.

dramatically different over the genes we tested. However, robustness as measured by the proportion of gene-causal-SNP cases with lower power than the Wald test was better with CLQ than with LDSelect. The plot of entire power values for all trait-causal SNP models over 1000 genes also indicated that the MLC test using CLQ is less likely than LDSelect to have substantially reduced power relative to the Wald test. Because the CLQ algorithm produces slightly more clusters than LDSelect for a given threshold, the degrees of freedom tends to be higher for the MLC test using CLQ than LDSelect, and in that matter CLQ has a disadvantage compared to LDSelect. However, the smaller sized clusters constructed

by CLQ may be advantageous because SNPs with opposing effects are less likely to occur in the same cluster.

For power comparisons using different threshold values, we found that CLQ with threshold values $c = 0.4\sim0.5$ usually produces the best average power among all clustering algorithm-threshold value combinations. In previous studies [8, 13], we suggested the threshold value 0.3~0.5 for r^2 (0.5~0.7 for $|r| \leq c$) to achieve optimum power using LDSelect algorithm, which has been validated by the results of this study. We suggest the threshold value of $c = 0.4\sim0.5$ to be used for CLQ algorithm based on the results of this study. However, a dynamically determined threshold value after evaluating the LD structure might be more appropriate for MLC tests, and being able to choose nonarbitrary threshold values is more attractive to researchers applying the method.

We applied the CLQ algorithm to a prespecified gene unit for gene-based analysis, but it could be applied similarly to intergenic regions, exomes, and promoter regions, that is, any regional units exhibiting some linkage disequilibrium between SNPs. If these regions include too many SNPs (e.g., more than 100), it is unreasonable to apply MLC tests based on joint regression models unless the sample size is extremely large. In that case, it may be desirable to break up the region into several LD blocks. Another approach would be to apply variable selection techniques such as penalized regressions [25, 26] and construct a MLC-type test with the resulting models.

5. Conclusions

In summary, we observed that CLQ and LDSelect produce identical clusters about half the time, and in the remaining cases, CLQ usually produces more clusters of smaller size. On average, MLC test power using CLQ is similar to that using LDSelect. The MLC test using CLQ shows better robustness to the detrimental effects of opposing SNP associations within the same cluster. Therefore, the CLQ algorithm is a promising approach for preanalysis clustering of SNPs for multimarker methods such as the MLC test.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) Grant NRF-2012R1A1A3012428 and the Canadian Institutes of Health Research (CIHR) operating Grant MOP-84287.

References

[1] D. J. Schaid, S. K. McDonnell, S. J. Hebbring, J. M. Cunningham, and S. N. Thibodeau, "Nonparametric tests of association of multiple genes with human disease," *The American Journal of Human Genetics*, vol. 76, no. 5, pp. 780–793, 2005.

[2] T. Wang and R. C. Elston, "Improved power by use of a weighted score test for linkage disequilibrium mapping," *The American Journal of Human Genetics*, vol. 80, no. 2, pp. 353–360, 2007.

[3] 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.

[4] J. L. Asimit, Y. Yoo, D. Waggott, L. Sun, and S. B. Bull, "Region-based analysis in genome-wide association study of Framingham Heart Study blood lipid phenotypes," *BMC Proceedings*, vol. 3, supplement 7, article S127, 2009.

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

[6] J. Z. Liu, A. F. McRae, D. R. Nyholt et al., "A versatile gene-based test for genome-wide association studies," *American Journal of Human Genetics*, vol. 87, no. 1, pp. 139–145, 2010.

[7] X. Wang, N. J. Morris, X. Zhu, and R. C. Elston, "A variance component based multi-marker association test using family and unrelated data," *BMC Genetics*, vol. 14, article 17, 2013.

[8] Y. J. Yoo, L. Sun, and S. B. Bull, "Gene-based multiple regression association testing for combined examination of common and low frequency variants in quantitative trait analysis," *Frontiers in Genetics*, vol. 4, article 233, 2013.

[9] B. M. Neale and P. C. Sham, "The future of association studies: gene-based analysis and replication," *The American Journal of Human Genetics*, vol. 75, no. 3, pp. 353–362, 2004.

[10] D. Curtis, "A rapid method for combined analysis of common and rare variants at the level of a region, gene, or pathway," *Advances and Applications in Bioinformatics and Chemistry*, vol. 5, pp. 1–9, 2012.

[11] 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," *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.

[12] D. Clayton, J. Chapman, and J. Cooper, "Use of unphased multilocus genotype data in indirect association studies," *Genetic Epidemiology*, vol. 27, no. 4, pp. 415–428, 2004.

[13] Y. J. Yoo, L. Sun, J. Poirier, and S. B. Bull, "Multi-bin multivariant tests for gene-based linear regression analysis of genetic association," Tech. Rep., Department of Statistical Sciences, University of Toronto, 2014, http://www.utstat.toronto.edu/wordpress/?page_id=514.

[14] B. M. Neale, M. A. Rivas, B. F. Voight et al., "Testing for an unusual distribution of rare variants," *PLoS Genetics*, vol. 7, no. 3, Article ID e1001322, 2011.

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

[16] 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," *The American Journal of Human Genetics*, vol. 92, no. 6, pp. 841–853, 2013.

[17] A. Derkach, J. F. Lawless, and L. Sun, "Pooled association tests for rare genetic variants: a review and some new results," *Statistical Science*, vol. 29, no. 2, pp. 302–321, 2014.

[18] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS Genetics*, vol. 5, no. 2, Article ID e1000384, 2009.

[19] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *The American Journal of Human Genetics*, vol. 74, no. 1, pp. 106–120, 2004.

[20] C. Bron and J. Kerbosch, "Algorithm 457: finding all cliques of an undirected graph," *Communications of the ACM*, vol. 16, no. 9, pp. 575–577, 1973.

[21] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal: Complex Systems*, p. 1695, 2006.

[22] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[23] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[24] C. J. Willer, E. M. Schmidt, S. Sengupta et al., "Discovery and refinement of loci associated with lipid levels," *Nature Genetics*, vol. 45, no. 11, pp. 1274–1283, 2013.

[25] K. L. Ayers and H. J. Cordell, "Identification of grouped rare and common variants via penalized logistic regression," *Genetic Epidemiology*, vol. 37, no. 6, pp. 592–602, 2013.

[26] K. L. Ayers and H. J. Cordell, "SNP Selection in genome-wide and candidate gene studies via penalized logistic regression," *Genetic Epidemiology*, vol. 34, no. 8, pp. 879–891, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 462549, 13 pages
http://dx.doi.org/10.1155/2015/462549

Research Article

Identifying and Assessing Interesting Subgroups in a Heterogeneous Population

Woojoo Lee,[1,2] Andrey Alexeyenko,[3] Maria Pernemalm,[4] Justine Guegan,[5] Philippe Dessen,[5] Vladimir Lazar,[5] Janne Lehtiö,[4] and Yudi Pawitan[1]

[1]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden*
[2]*Department of Statistics, Inha University, Incheon 402-751, Republic of Korea*
[3]*Department of Microbiology, Tumour and Cell Biology, Bioinformatics Infrastructure for Life Sciences, Science for Life Laboratory, Karolinska Institutet, 17177 Stockholm, Sweden*
[4]*Department of Oncology and Pathology, Science for Life Laboratory, Karolinska Institutet, 17121 Solna, Sweden*
[5]*Genomics, Institut Gustave Roussy, F-94805 Villejuif, France*

Correspondence should be addressed to Yudi Pawitan; Yudi.Pawitan@ki.se

Received 10 November 2014; Revised 1 March 2015; Accepted 3 March 2015

Academic Editor: Kristel van Steen

Biological heterogeneity is common in many diseases and it is often the reason for therapeutic failures. Thus, there is great interest in classifying a disease into subtypes that have clinical significance in terms of prognosis or therapy response. One of the most popular methods to uncover unrecognized subtypes is cluster analysis. However, classical clustering methods such as k-means clustering or hierarchical clustering are not guaranteed to produce clinically interesting subtypes. This could be because the main statistical variability—the basis of cluster generation—is dominated by genes not associated with the clinical phenotype of interest. Furthermore, a strong prognostic factor might be relevant for a certain subgroup but not for the whole population; thus an analysis of the whole sample may not reveal this prognostic factor. To address these problems we investigate methods to identify and assess clinically interesting subgroups in a heterogeneous population. The identification step uses a clustering algorithm and to assess significance we use a false discovery rate- (FDR-) based measure. Under the heterogeneity condition the standard FDR estimate is shown to overestimate the true FDR value, but this is remedied by an improved FDR estimation procedure. As illustrations, two real data examples from gene expression studies of lung cancer are provided.

1. Introduction

Biological heterogeneity is common in many diseases; heterogeneity complicates clinical management, as it is often the reason for prognostic and therapeutic failures. Thus, there have been many attempts to classify a disease into subtypes with anticipation that different subgroups are associated with different clinical significance in terms of prognosis or therapy response (e.g., [1, 2]). A significant progress in designing efficient specific treatments can be achieved if novel clinically relevant subtypes are found.

One of the most popular methods for finding unrecognized subtypes is cluster analysis. However, classical clustering methods such as k-means clustering or hierarchical clustering are not guaranteed to produce clinically interesting subtypes because the main statistical variability could be dominated by genes not associated with interesting clinical phenotypes. Furthermore, it could be that prognostic factors shared within a subgroup do not have any important role in other subgroups. Thus, the association between prognostic factors and a clinical phenotype is attenuated and not detectable in the whole population. To address these problems, we extend the standard clustering algorithm to find interesting subgroups in the sense that within the subgroup we can find factors (in this paper: genes) strongly associated with the clinical phenotype. This idea can perhaps be illustrated more clearly as follows: suppose that Y is an outcome (e.g., relapse) and X is a randomized treatment;

it is common to search for a subgroup for which the treatment effect is largest. In effect we are searching for factors Z that have significant interactions with X, such that a subgroup defined by Z will have a large treatment effect on Y. A unique point in our current application is that both X and Z are given by the same set of gene expression data. Also, we allow complex subgroups to be discovered by a clustering method, which makes the process distinct from the standard interaction analysis.

Given a set of gene expression matrix, our goal of cluster analysis is to group patients and genes into subgroups that convey biological or clinical significance. This task can be translated to the biclustering problem. Biclustering methods attempt to simultaneously cluster both patients and genes with the goal of finding subsets of rows and columns in the expression matrix. Cheng and Church [3] firstly introduced biclustering to gene expression analysis. For reviewing the details of biclustering algorithms, see [4]. As Nowak and Tibshirani [5] noticed, however, most of biclustering algorithms tend to be dominated by groups of highly differentially expressed (DE) genes that may not be relevant to the biological process in question. In other words, irrelevant genes with strong signal can mask genes of highest biological relevance. Furthermore, iterative optimization methods adopted in biclustering algorithms depend on initial conditions. To overcome these limitations, we develop an extensive clustering search algorithm to find molecular subtypes (CAMS) based on clustering of patients with partially similar mRNA profile. CAMS is able to uncover the structures arising from relevant genes that may not be highly expressed but moderately expressed within each subtype.

CAMS produces many subtypes. For each subtype, t-statistics comparing two distinct phenotype groups (e.g., relapsed/not relapsed) are computed for whole genes and false discovery rate (FDR) estimate is used to correct for multiple comparisons. The number of genes having small FDR estimates (say, less than 0.1) is the basis for assessing the importance of the subtypes. In real data analysis, however, it is a common occurrence in heterogeneous populations that P value distributions of the two-sample t-statistics show substantial shortage of small values compared to the uniform distribution [6]. If we ignore this effect we would miss potential discoveries by overestimating FDR. Since subtypes produced by CAMS still can be heterogeneous, it is crucial to study how the molecular heterogeneity of distinct subtypes affects the FDR estimate. In this paper, we introduce unobserved group (or latent group) variables into a simple model for gene expression and see how the heterogeneity induced by the unobserved group leads to the depletion of small P values even when there are many significant signatures. Thus, without considering this underlying heterogeneity, the use of standard FDR estimate might hide promising discoveries. To resolve this problem, we develop an improved FDR estimation procedure to address the heterogeneity in a dataset.

In estimating FDR, the use of correct null density function is critical. Efron [7] considered three issues that substantially affect the null density estimate in computing FDR: (1) a large proportion of genuine but uninterestingly small effects, (2) hidden correlations, and (3) unobserved covariates. Many researchers have studied how they affect the standard FDR estimate [7–9]. In particular, possible connections between unobserved covariates and FDR have been explored in [6, 10]. Leek and Storey [6] showed numerically that the small P values range from being inflated to depleted depending on the configuration of the unobserved covariates. They developed the so-called surrogate variable analysis (SVA) for capturing heterogeneity induced by the unobserved covariates and studied how SVA affects FDR estimate. Stegle et al. [10] considered a Bayesian method to account for hidden confounding variation in expression quantitative trait loci (QTLs) and showed that the method found additional expression QTLs in real datasets. However, their approaches were suggested to study the attenuated relationship by heterogeneity between a measured variable of interest and clinical outcomes, while we focus on finding submerged subtypes by heterogeneity. The novel contributions of this paper are (1) to explain how the heterogeneity induced by unobserved group leads to the depletion of small P values analytically, (2) to analyze the bias of standard FDR estimates under the heterogeneity, and (3) to develop an improved FDR estimation procedure. With these in mind a FDR-based measure is considered to assess findings from a novel clustering procedure. This is illustrated using two datasets on lung cancer patients.

The rest of this paper is organized as follows. In Section 2, we describe the implementation details of CAMS. A brief review of notations and a standard FDR estimation method are given in Section 3, and it is analytically shown that the hidden subgroup in the population can induce a bias of standard FDR estimate in Section 4. We propose a FDR estimation procedure resolving the bias problem and show how to assess clustering results from CAMS with it in Sections 5 and 6. Section 7 includes two real data applications and is followed by concluding remarks.

2. Clustering Algorithm for Finding Molecular Subtypes

Consider a set of gene expression profiles from a group of cancer patients. The premise behind CAMS is that the novel molecular information on cancer heterogeneity is hidden in the gene expression profiles. To uncover the heterogeneity, CAMS implements a two-dimensional clustering "patients versus genes" extensively. The full algorithm is given in Algorithm 1.

We first explain the clustering steps of CAMS graphically in Figures 1(a) and 1(b). In the two figures, a set of gene expression profiles as a matrix with rows corresponding to genes and columns corresponding to patients is graphically represented. For illustrative purposes, we designed the following simple model.

(i) It has two observed groups: for example, relapse yes (RY) and relapse no (RN) groups.

(ii) It has two unobserved groups: the first two columns correspond to molecular subtype 1 (MS1) and the remaining two columns correspond to molecular

```
while r ∈ {1, 2, ..., n_perm} do
  Shuffle the genes
  Partition the genes into S disjoint subsets
  for i = 1 to S do
    Perform hierarchical clustering on the ith subset of the genes
    for the number of clusters (c) ∈ C do
      Cut the dendrogram from the hierarchical clustering to yield c clusters.
      for j = 1 to c do
        Take jth cluster (a subtype identifier)
        Run hierarchical clustering on patients using only the genes
        in jth cluster (Assume that this step yields K clusters).
        for k = 1 to K do
          Perform two-sample t-tests (e.g. relapse yes vs. no)
          with the individuals in kth cluster and the whole genes.
          Fit the null distribution of the two sample t-statistics
          with known functional forms.
          if
        (Normal approximation for the null distribution is acceptable) then
            Compute the FDR estimate based on the normal approximation.
          end if
          if
        (Normal approximation for the null distribution is not acceptable) then
            Compute the FDR estimate based on K permutations.
          end if
          Compute our proposed FDR-based measure (N01)
          to assess the resulting cluster.
          For a given subtype, compute the P-value of N01 by permuting group labels.
        end for
      end for
    end for
  end for
  r ← r + 1
end while
```

Algorithm 1: CAMS.

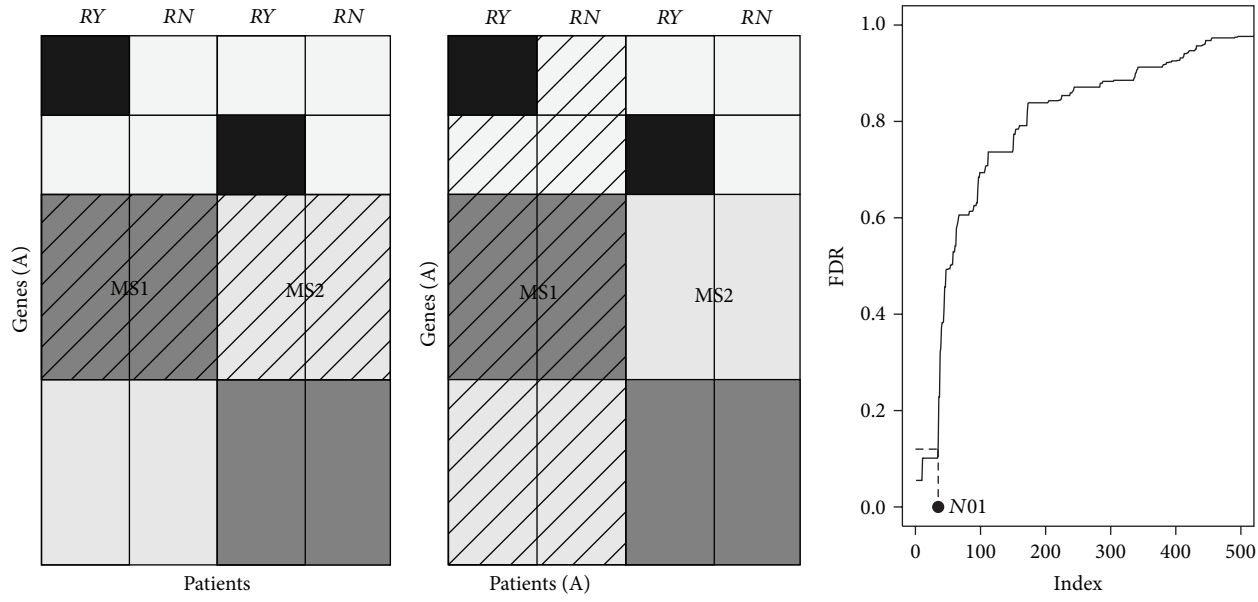
subtype 2 (MS2). This information is unknown to the researchers.

(iii) Some genes (marked in black) affect relapse within a MS.

The two key clustering steps of CAMS are as follows. Step I is clustering of genes. This step identifies several sets of genes having similar profiles across the patients. For example, in Figure 1(a), gene-set A (shaded region) is grouped and this will be used as a subtype identifier in next step. Step II is clustering of patients using gene-set A only. This step produces a subgroup of patients (individuals belonging to the shaded region of Figure 1(b)) with a common expression profile for gene-set A. Note that this subgroup is homogeneous in terms of the identified set of genes from the first step but can show distinct expression profiles between *RY* and *RN* on the other set of genes (e.g., genes marked in black). Thus, we hope that, within the subgroup of patients, good prognostic models can be constructed.

Technical description for CAMS is given as follows. In Step I, the set of gene probes on the microarray chip is grouped via hierarchical clustering. This is implemented using *hclust* in *R*. All the hierarchical clustering in this paper uses complete linkage method and Euclidean metric. This hierarchical clustering procedure is applied to disjoint subsets (S) of m all available gene probes due to computational limit (e.g., $m = 41000$ gene probes in the lung cancer dataset). For example, if $S = 10$, our procedure makes 10 disjoint subsets of gene probes and each subset has $m/10$ gene probes sampled from the whole list. Then the whole list is systematically covered by applying the clustering to S subsets sequentially. To allow various groupings of gene probes under different environments, we shuffle the whole list of gene probes several times. The number of clusters (C) from each subset S varies on a vector of fixed numbers. For example, if $C = (2, 3, 4, 5, 6, 7, 8, 9, 10)$, then 9 different cluster analysis results are considered in the downstream analysis. Thus each gene probe could participate in different clustering solutions, from very large (>500 probes) to small sets. These clustering results can be used as subtype identifiers in next step.

In Step II, the same hierarchical clustering method is applied to cluster the patients by using each subtype identifier separately. Then the dendrogram is cut at the highest level where the clusters contain more patients than the threshold. Each subset of patients is treated as a candidate subtype.



(a) Step I (b) Step II (c) Step III

FIGURE 1: (a) Step I is clustering of genes. Genes (A) (shaded region) are grouped and will be used as a subtype identifier in the downstream analysis. (b) Step II is clustering of patients using Genes (A) (i.e., a gene-set obtained from Step I). Here, MS1 (a set of patients, shaded regions) is obtained as a subtype. (c) Step III shows how $N01$ is obtained from the FDR curve. We count the number of genes having FDR < 0.1. We repeat implementing (a), (b), and (c) across different clustering results extensively. Thus, no shaded columns in (b) will be covered subsequently.

When the clustering steps of CAMS are performed, only some of found cancer subtypes would be true discoveries. To assess whether subtypes are promising or not, t-statistics comparing two distinct phenotype groups (e.g., relapsed/not relapsed) within each subtype are computed for whole genes and the number of genes having small FDR estimates (say, less than 0.1) is calculated based on the P values of the t-statistics. However, the effect of the molecular heterogeneity on this assessment has not been explored in detail. To deal with this issue, we first review a standard FDR estimation method below.

3. Notation and Standard FDR Estimation

In this section some basic notations are introduced to give a formal definition of FDR. For clarity and simplicity, we will limit our discussion to the most common problem of finding differentially expressed (DE) genes between two biological conditions. Let z be a certain statistic to compare the mean log-expression level. The distribution of observed statistics z follows a mixture model

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \tag{1}$$

where π_0 is the proportion of truly nondifferentially expressed (non-DE) genes and $f_0(z)$ and $f_1(z)$ are the density functions of z for non-DE and DE genes, respectively.

Suppose we test m genes with corresponding statistics $z_1, \ldots, z_m$. Let $P_1, \ldots, P_m$ be the ordered P values from m test statistics. For a fixed critical value c, we define the number of non-DE genes declared DE and the number of genes declared DE as

$$V(c) = \sum_i I_{(P_i \le c, i \in \text{Null})},$$
$$R(c) = \sum_i I_{(P_i \le c)}, \tag{2}$$

where $I(\cdot)$ is the indicator function. Then, the false discovery proportion (FDP) is defined as

$$\text{FDP}(c) = \frac{V(c)}{R(c)}, \tag{3}$$

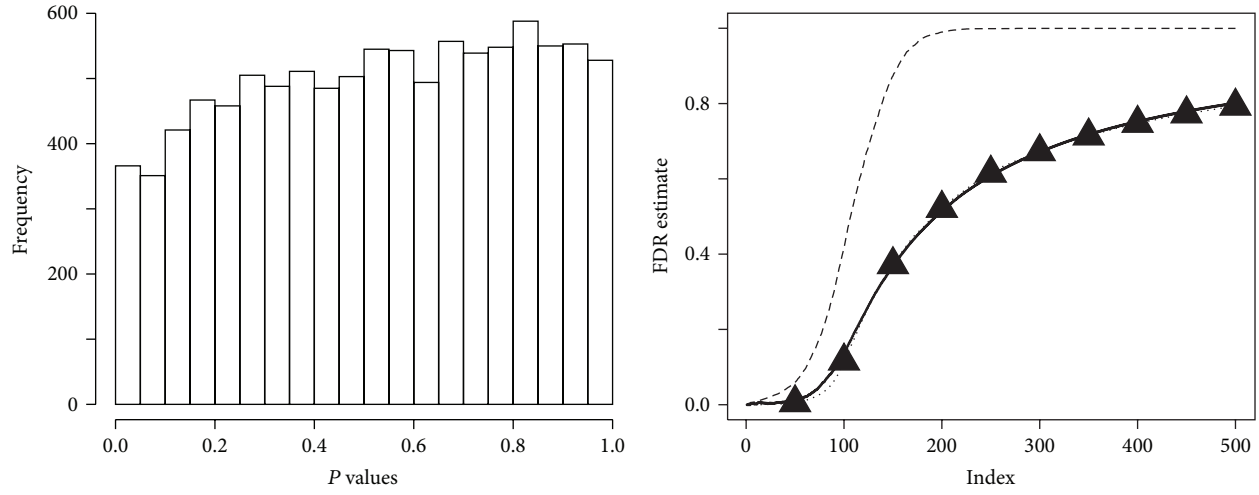
except in the case of $R(c) = 0$, in which case we just set $\text{FDP}(c) = 0$. The FDP is random proportion of false discoveries among the genes declared to be DE. The standard FDR is the marginal average of the FDP; namely, $\text{FDR}(c) = E(\text{FDP}(c))$.

The standard estimate of FDR [8, 11] as a function of the ordered P values is given by

$$\widehat{\text{FDR}}(P_k) = \frac{m\hat{\pi}_0 P_k}{k}. \tag{4}$$

Monotonicity is imposed by taking the cumulative minimum over $\widehat{\text{FDR}}(P_i)$ $(i = k, \ldots, m)$. A common used formula for $\hat{\pi}_0$ is

$$\hat{\pi}_0 = \frac{(\text{Number of } P \text{ values} > \lambda)}{(m(1 - \lambda))} \tag{5}$$



(a) Histogram of P values

(b) Comparison of FDR estimates

FIGURE 2: (a) P value $= P(|T| \geq |t_{\text{obs}}|)$, where T is a generic two-sample t-statistic and t_{obs} is an observed t-statistic and (b) average from 50 simulations: true false discovery proportion (FDP) (solid), standard estimate (dashed), and proposed procedure (dotted). The dotted line coincides with the solid, so it is additionally marked with triangles.

for a certain choice of λ [11]. Simple choices of λ such as 0.5 or 0.75 are often used. Note that this standard estimation procedure does not consider the heterogeneity in population.

4. A Bias of the Standard FDR Estimate

Latent variables have been introduced for various purposes in multiple testing framework. Friguet et al. [12] and Leek and Storey [13] considered them as a source of dependence among genes. In this paper we introduce latent variables as a source of heterogeneity and design a latent group model leading to a depleted P value distribution near 0. With practical applications in mind, we will adopt terminologies from two-sample microarray studies for cancer. Our toy model is already graphically represented in Section 2. There are two unobserved groups (molecular subtypes 1 (MS1) and 2 (MS2)) and two observed groups (relapse yes (RY) and relapse no (RN)). Genes affecting relapse within a MS are marked black and genes identifying MS are marked dark gray. More details to generate Figure 1 are as follows.

(i) For most genes, we choose one MS randomly with probability 0.5 and generate background effects from $N(\mu/2, 1)$. For other MS, we generate background effects from $N(0, 1)$. These genes are used to define specific molecular subtypes (MS1 and MS2) and are undiscriminating for the two observed groups (RY and RN).

(ii) Some genes affect relapse within a MS. After choosing one MS with probability 0.5, we generate background effects from $N(\mu/2, 1)$. Then, we add signal effects generated from $N(\mu_0/2, 1)$ for RY and $N(-\mu_0/2, 1)$ for RN, respectively. For other MS, we generate background effects from $N(0, 1)$.

Consider genes defining MS and highly expressed in MS1. Then, we have the following ANOVA representation:

$$Y_{ij} = \left(\frac{\mu}{2}\right) I_{(j\in \text{MS1})} + \varepsilon_{ij}, \tag{6}$$

where i is the index for gene, j is the index for patient, and $\varepsilon_{ij} \sim N(0, 1)$. For relapse-related genes within MS1, we have the following ANOVA representation:

$$Y_{ij} = \left(\frac{\mu}{2}\right) I_{(j\in \text{MS1})} + \left(\frac{\mu_0}{2}\right) I_{(j\in RY)} + \left(-\frac{\mu_0}{2}\right) I_{(j\in RN)} + \varepsilon^*_{ij}, \tag{7}$$

where $\varepsilon^*_{ij} \sim N(0, 2)$. In contrast to our model, Efron [7] considered the following model:

$$Y_{ij} = \left(\frac{\mu_{0i}}{2}\right) I_{(j\in RY)} - \left(\frac{\mu_{0i}}{2}\right) I_{(j\in RN)} + \varepsilon_{ij}, \tag{8}$$

where $\mu_{0i} \sim N(0, \sigma^2)$. Note that this model does not consider unobserved group, and as [7] pointed out, this model can lead to only a dilated null distribution that explains inflation of small P values (i.e., false positives). In Figure 2(a), however, our latent group model shows the depletion of small P values. Note that (6) dominates the overall shape of the P value distribution because it has high proportion in the model.

We now see how the unobserved group in the population induces a bias of standard FDR estimate. As a first step, we compute two-sample t-statistic to compare RY and RN. In RY, there are n_y patients, where the half are from MS1 and the other from MS2. In RN, there are n_n patients and it has the same structure. Thus, RY and RN groups consist of two normal distributions having different means. Consider the genes following (6). Let $\overline{RY}_i = \sum_j Y_{ij} 1_{(j\in RY)}/n_y$ and

$\overline{RN}_i = \sum_j Y_{ij} 1_{(j\in RN)}/n_n$. The t-statistic to test the null hypothesis (non-DE) is

$$z_i = \frac{\left(\overline{RY}_i - \overline{RN}_i\right)}{\left(\widehat{\sigma}_i \sqrt{1/n_y + 1/n_n}\right)}, \tag{9}$$

where

$$\widehat{\sigma}_i = \sqrt{\frac{\left(\sum_j \left(Y_{ij} - \overline{RY}_i\right)^2 1_{(j\in RY)} + \sum_i \left(Y_{ij} - \overline{RN}_i\right)^2 1_{(j\in RN)}\right)}{\left(n_y + n_n - 2\right)}}. \tag{10}$$

Note that, for large n_y and n_n, we have

$$\widehat{\sigma}_i \longrightarrow_p \sqrt{1 + \frac{\mu^2}{16}}, \tag{11}$$

because

$$\begin{aligned}
&\frac{\sum_j \left(Y_{ij} - \overline{RY}_i\right)^2 1_{(j\in RY)}}{n_y} \\
&= \frac{\sum_j \left(Y_{ij} - \mu/4\right)^2 1_{(j\in RY)}}{n_y} + o_p(1) \\
&= \frac{\sum_{(j\in RY \bigcap \mathrm{MS1})} \left(Y_{ij} - \mu/4\right)^2}{n_y} \\
&\quad + \frac{\sum_{(j\in RY \bigcap \mathrm{MS2})} \left(Y_{ij} - \mu/4\right)^2}{n_y} + o_p(1) \\
&\longrightarrow_p 0.5\left(1 + \frac{\mu^2}{16}\right) + 0.5\left(1 + \frac{\mu^2}{16}\right) \\
&= \left(1 + \frac{\mu^2}{16}\right), \\
&\frac{\sum_j \left(Y_{ij} - \overline{RN}_i\right)^2 1_{(j\in RN)}}{n_n} \\
&= \frac{\sum_j \left(Y_{ij} - \mu/4\right)^2 1_{(j\in RN)}}{n_n} + o_p(1) \\
&\longrightarrow_p \left(1 + \frac{\mu^2}{16}\right).
\end{aligned} \tag{12}$$

Meanwhile, the numerator in (9) is

$$\begin{aligned}
&\sqrt{n}\left(\overline{RY}_i - \overline{RN}_i\right) \\
&= \sqrt{n}\left(\frac{\sum_j Y_{ij} 1_{(j\in RY)}}{n_y} - \frac{\sum_j Y_{ij} 1_{(j\in RN)}}{n_n}\right) \\
&= \sqrt{n}\left(\frac{\sum_{(j\in RY \bigcap \mathrm{MS1})} \left(Y_{ij} - \mu/2\right)}{n_y} + \frac{\sum_{(j\in RY \bigcap \mathrm{MS2})} Y_{ij}}{n_y}\right. \\
&\qquad \left. - \frac{\sum_{(j\in RN \bigcap \mathrm{MS1})} \left(Y_{ij} - \mu/2\right)}{n_n} - \frac{\sum_{(j\in RN \bigcap \mathrm{MS2})} Y_{ij}}{n_n}\right) \\
&\longrightarrow_d N(0, 4).
\end{aligned} \tag{13}$$

Thus, we have for large n

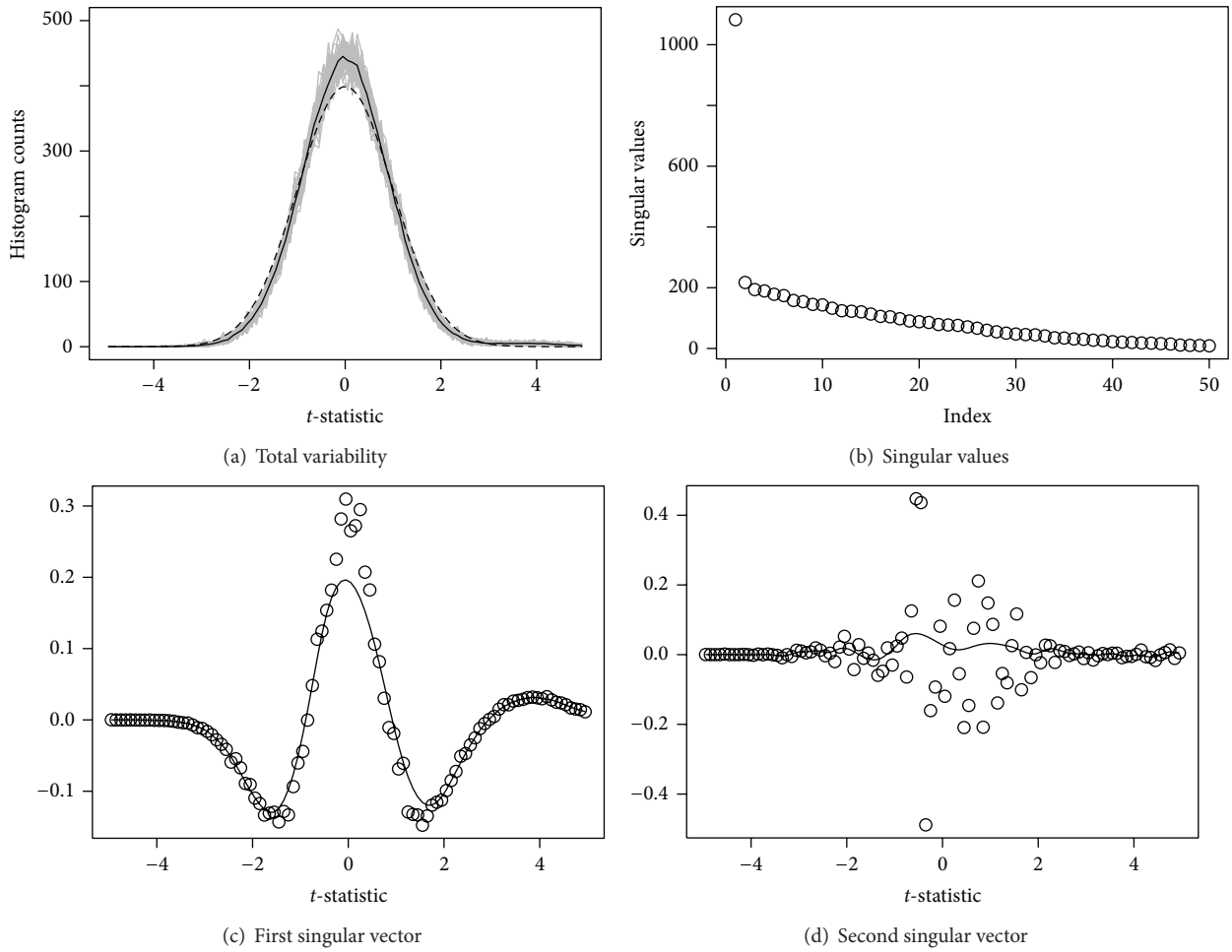
$$z_i \longrightarrow_d N\left(0, \frac{1}{\left(1 + \mu^2/16\right)}\right). \tag{14}$$

Since $1 + \mu^2/16 > 1$ for any $\mu \neq 0$, the use of standard Gaussian distribution for z_i leads to inflated P values. Thus, in (4), $\widehat{\pi}_0$ is overestimated and $R(c)$ is smaller than it should be. Subsequently, (4) overestimates FDR. If the strength of background signal μ becomes larger, the degree of depletion of small P values becomes more severe because (14) will be more concentrated at 0 as μ increases. Consequently, the heterogeneity induced by the unobserved group makes the t-statistics conservative and leads to upward bias of standard FDR estimate as shown in Figure 2(b). In our simulation, we use 10,000 genes and 60 patients, with 30 belonging to each MS. The proportion of genes defining specific MS is 0.99. Within each MS, the number of RY and RN is assumed to be same for simplicity and we use $\mu = 2$ and $\mu_0 = 3$.

5. Proposed FDR Estimation Procedure

While performing CAMS, we want to assess whether clustering results are informative or not with respect to a measure based on FDR. Thus, in computing FDR estimate, the population heterogeneity should be addressed properly. Furthermore, when many datasets are considered simultaneously, it is desirable to have a fast and stable algorithm to compute FDR estimate. Reflecting these aspects, we propose a new FDR estimation procedure.

Our starting point is Pawitan et al.'s FDR estimation procedure [9] because it is computationally flexible to accommodate new changes. A similar permutation-based approach to deal with the dependence in computing FDR estimates was developed by [14]. Pawitan et al. [9] explored the variation pattern of the null distribution of test statistics using the singular value decomposition (SVD) when there are correlations between genes. To check the validity of the SVD analysis in our problem, it is needed to confirm whether the main variation pattern of permutation distribution can represent that of sampling distribution.



(a) Total variability (b) Singular values

(c) First singular vector (d) Second singular vector

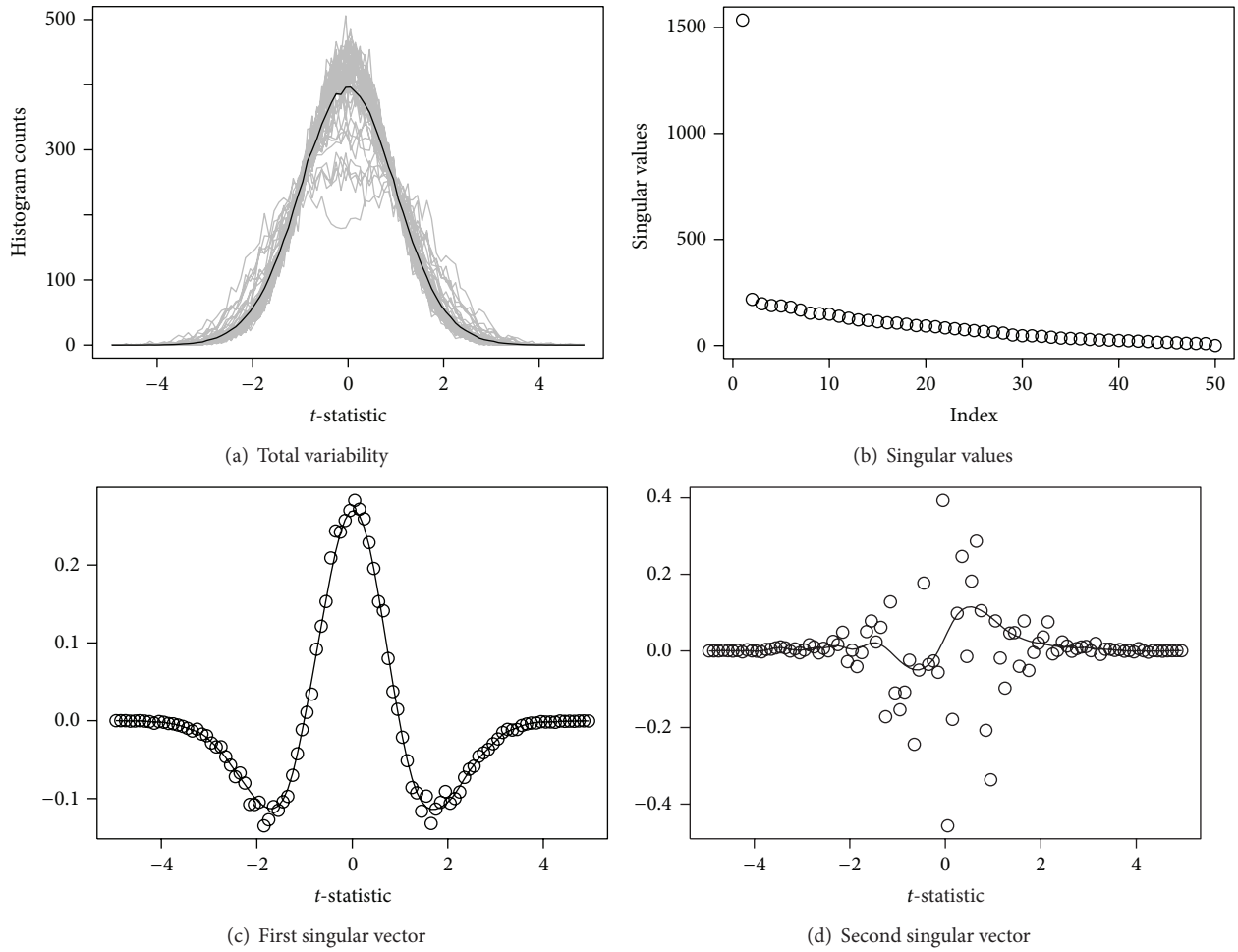
FIGURE 3: (a) Each simulation contributes a single gray line. The solid black line is the average of 50 simulations, and dashed line is the expected histogram-count vector from $N(0, 1)$. (b) shows singular values from the singular value decomposition (SVD) of Y, and the dots in (c) and (d) are the components of the singular vectors generated by the SVD, and the solid lines are robust smoothing curves.

5.1. The Validity of SVD Analysis. We firstly demonstrate the variation pattern of the sampling distributions from the latent group model through the SVD analysis. We partition the range of the observed statistics into B equispaced bins with width Δ. Let the histogram-count vector $\mathbf{y} = (y_1, \ldots, y_B)$ be the number of statistics that fall into each bin. Each simulation contributes a single count vector $\mathbf{y_i}$. Let $\boldsymbol{\eta}$ be the expected histogram-count vector from standard Gaussian distribution and the $B \times K$ matrix Y the matrix of centered count vectors $\mathbf{y_i} - \boldsymbol{\eta}$. K is the number of simulations and 50 is used in our example.

Figure 3(a) shows the total variability of sampling distributions; the solid line is $\bar{\mathbf{y}}$ and the dashed line is $\boldsymbol{\eta}$. The solid line has higher peak and smaller width than the dashed line, so this is consistent with our analytical findings. To see the variability of $\mathbf{y_i} - \boldsymbol{\eta}$, we perform the SVD of Y. The variation is dominated by one large singular value, associated with the pattern seen in the plot of the first singular vector. A consequence of this pattern is that the sampling distribution tends to have a leptokurtic shape compared to the standard Gaussian distribution. Subsequent singular vectors do not have large contributions to the variation.

In practice, we cannot create real data as in simulation. To circumvent this problem, we use permutation to generate the null distribution, but we first check the variability pattern of the distributions from permutation. Let X be a microarray data matrix, let $\mathbf{g} = (g_1, \ldots, g_n)$ be the vector of group labels, and let g^* be a random rearrangement of g. With each permuted dataset (X, g^*), we compute test statistics. So each permutation contributes a single count vector $\mathbf{y_i^*}$. Let $\bar{\mathbf{y}}^*$ be the mean vector of $\mathbf{y_i^*}$ over K permutations and the $B \times K$ matrix Y^* the mean-corrected matrix of count vector $\mathbf{y_i^*}$. The SVD results of Y^* are reported in Figure 4.

Figure 4(a) shows the total variability of the distribution over permutations, and the solid line is the average of the permuted null distributions. In Figure 4(b), the first singular value is dominating others and Figure 4(c) shows that the pattern of the first singular vector from Y^* is very close to that from Y. This implies that the main variation of permuted distributions explains that of the sampling distributions well,



(a) Total variability

(b) Singular values

(c) First singular vector

(d) Second singular vector

Figure 4: (a) Each permutation contributes a single gray line. The solid black line is the average of 100 permutations. (b) shows singular values from the singular value decomposition (SVD) of Y^*, and the dots in (c) and (d) are the components of the singular vectors generated by the SVD, and the solid lines are robust smoothing curves.

so the SVD analysis for permuted data is valid under our latent group model.

Since we have the validity of the SVD analysis, Pawitan et al.'s method [9] can be adopted to correct the overestimation by unobserved group. We assume that the observed statistics z follow a mixture model (1). They suggested to fit

$$\mathbf{y} \sim \text{Poisson}\left(m\Delta f\left(z\right)\right), \tag{15}$$

where

$$f\left(z\right) = \pi_0\left(\phi_0\left(z\right) + b\phi_1\left(z\right)\right) + \left(1-\pi_0\right) f_1\left(z\right), \tag{16}$$

where $f_0(z) = \phi_0(z) + b\phi_1(z)$, $\phi_0(z)$ is the average of null distributions over permutations, and $\phi_1(z)$ is the first singular vector of Y^*. In this paper, the parameter b captures the variation of the null distribution due to the heterogeneity by unobserved group. The original computing procedure is given as follows.

(1) Perform K permutations of group labels. Each permuted dataset generates a histogram-count vector $\mathbf{y}^*$.

(2) Compute the predictor ϕ_0 from the average vector $\overline{\mathbf{y}}^*$ by scaling so that it integrates to 1.

(3) Construct a matrix Y^* from the $\mathbf{y}^*$s. Compute the predictor ϕ_1 from the smoothed first singular vector u_1.

(4) Since f_1 is unknown, the regression is performed in two steps. First, fit the reduced model $\mathbf{y} \sim \text{Poisson}(\mu = m\Delta f)$, where

$$f = \beta_0\phi_0 + \beta_1\phi_1, \tag{17}$$

and compute the residual vector $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$. Estimate f_1 by smoothing the residual vector $\mathbf{r}/m\Delta$ as a function of z.

(5) Fit the full model (16)

$$f = \beta_0\phi_0 + \beta_1\phi_1 + \beta_2 f_1, \tag{18}$$

and reestimate the full set of coefficients $(\beta_0, \beta_1, \beta_2)$.

The coefficient of ϕ_0 becomes the estimate for π_0. Given estimates of parameters, P values inflated by the heterogeneity are corrected by using the following definition:

$$P \text{ value} = \int_{|z|\geq|z_{\text{obs}}|} \widehat{f_0}(z)\, dz, \tag{19}$$

where $\widehat{f_0}(z)$ is the null density estimate corrected by the first singular vector. For the FDR estimate, we have

$$\widehat{\text{FDR}}(c) = \frac{m\widehat{\pi}_0 \int_{|z|>c} \widehat{f}_0(z)\, dz}{\sum_i I(|z_i| > c)}. \tag{20}$$

(Strictly speaking, this is an FDP estimate rather than an FDR estimate.) Simulation studies show that this estimate has a negligible bias (Figure 2(b)). It may be possible to improve the null density estimate further using the second singular vector in some cases, but we will not attempt this here.

5.2. Improved Algorithm for Many Datasets. CAMS generates many subtypes. Since not all the subtypes are meaningful, it is needed to assess each of them quickly. In particular, for the subtypes showing the depletion of small P values, it is desirable to apply our FDR procedure to address such depletions.

One measure to assess subtypes from CAMS is the number of genes having FDR $<$ c, where c is a suitably chosen small value; we use $c = 0.1$ in our examples and call this measure $N01$. Figure 1(c) shows that $N01$ is obtained by counting the number of genes with FDR < 0.1. To compute $N01$ for many datasets, the previous procedure becomes

(1) computationally intensive: the permutation step takes long time;

(2) unstable: some null density estimates have negative values.

To increase computational speed, we note that $\widehat{f_0}(z)$ from the SVD analysis is empirically well approximated by $N(0, (1 - b/\sqrt{2})^2)$ for $|b| < 0.2$, which can be checked before the permutation step. But this approximation does not seem to be reliable when $|b| > 0.2$. Furthermore, the null density estimates often have negative values when b is large and this leads to a numerical problem in estimating FDR. Thus, we propose a more stable algorithm to find good approximation to the null density. The main idea is to pick up a few vectors $\mathbf{y}_i^*$ that are closest to the histogram counts of the observed test statistics $\mathbf{y}$ with respect to a certain metric. To emphasize goodness of fit at the center of the distribution, we use

$$\text{Dist}(\mathbf{y}, \mathbf{y}_i^*) = (\mathbf{y} - \mathbf{y}_i^*)^T W_{\mathbf{y}} (\mathbf{y} - \mathbf{y}_i^*), \tag{21}$$

where $W_{\mathbf{y}} = \text{Diag}(\mathbf{y})$ as a distance measure. This distance measure gives larger weights to the central part of the histogram. We find top 5 curves that minimize (21) and use their average as $\widehat{f_0}(z)$. For simplicity we estimate π_0 with (5). The resulting procedure thus becomes as follows.

(i) Before the permutation step, approximate $\widehat{f_0}(z)$ by $N(0, (1 - b/\sqrt{2})^2)$ and obtain ϕ_0 and ϕ_1 using known functional forms [15]. After fitting (16) as described in Steps 4 and 5 in the previous section, check whether $|b| < 0.2$ or not. If $|b| < 0.2$, compute FDR estimate.

(ii) In the case of $|b| > 0.2$, perform K permutations and compute (21) for each permuted dataset. Find the top 5 curves that minimize (21) and take the average as the null density estimate. Estimate π_0 using formula (5).

6. Assessing the Clustering Results from CAMS

CAMS can generate a practically unlimited number of candidate subtypes by permuting the gene probes for doing extensive search. If a subtype is depleted in small P values, it is desirable to assess it with $N01$. To see the proportion of subtypes requiring such assessment, we define the ratio of low P value areas as

$$\text{Ratio}(\lambda) = \frac{\sum_i^m 1_{(P_i \leq \lambda)}}{m\lambda}, \tag{22}$$

where $\lambda = 0.2$ is used in practice. The denominator corresponds to the expected number of P values less than λ when the null hypothesis holds. We regard Ratio < 1 as indicating the targeted situation (the depletion of small P values). When the whole set of patients shows the depletion, we often observe high proportion of potential subtypes with Ratio < 1, so it is safe to use $N01$ as a default assessment measure.

We provide an implementation of the proposed method as an *R* package at http://fafner.meb.ki.se/personal/yudpaw/. Two necessary inputs for the implementation are gene expression data matrix and corresponding group vector (a clinical outcome such as disease outcome, e.g., relapse indicator). To enable further analysis when there is auxiliary information such as survival time, the software stores the following results:

(i) genes defining a cancer subtype,

(ii) patient IDs that belong to a subtype,

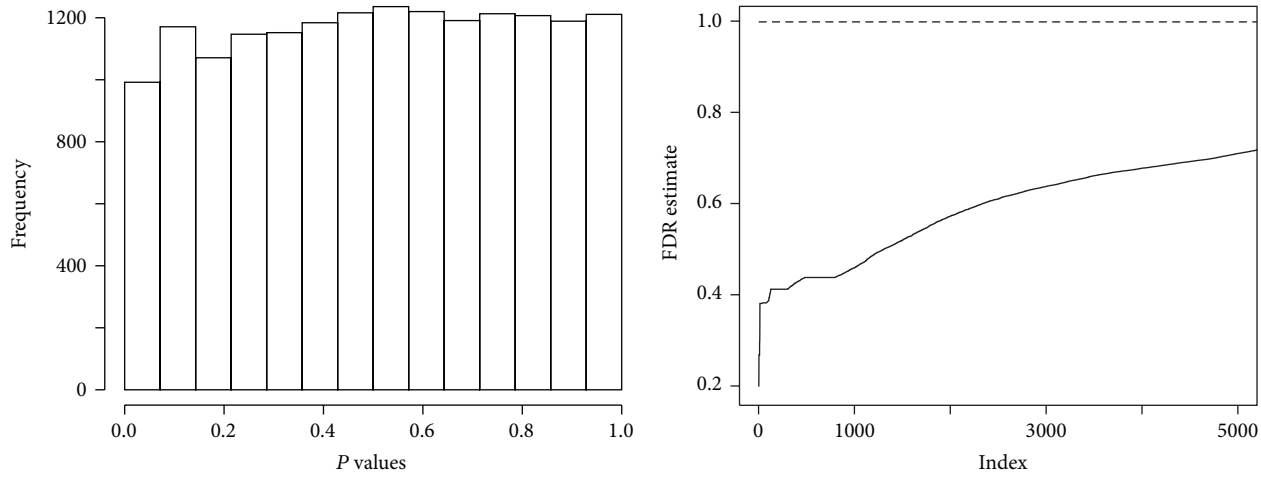
(iii) $N01$ and respective P value.

Note that we may have high $N01$ by chance because several optimization procedures (e.g., the biclustering procedure) are performed before computing $N01$. To address this point, we randomly permute group labels of each subtype N_p times and compute $N01$ based on the permuted data ($N01_{\text{perm}}$).

Then, we compute a standardized statistic of $N01$ for ith subtype:

$$z_i = \frac{N01_i - \overline{N01_i}}{s_i}, \tag{23}$$

where $\overline{\text{N01}_i}$ and s_i are the mean and standard deviation of $N01_i$ and $N01_{\text{perm}}$'s. $N_p = 50$ is used in practice. Likewise, we standardize $N01_{\text{perm}}$. This standardization enables us to have precise estimate for P value and reasonable resolution for estimating FDR. After stacking all the standardized statistics



(a) Histogram of P values (b) FDR

FIGURE 5: (a) P value distribution of two-sample t-statistics for detecting differentially expressed genes from a lung cancer data comparing relapse versus no relapse and (b) the corresponding false discovery rate (FDR) estimate. In (b), the dashed line is the standard FDR estimate and the solid line is from our proposed procedure. The x-axis in (b) denotes the ranking of genes where higher ranking corresponds to higher statistical significance.

in a vector z_{perm}, the P value of $N01$ for ith subtype is defined as

$$P\text{-value}_i = \frac{\sum_k I\left(z_i \le z_{\text{perm},k}\right)}{K}, \tag{24}$$

where K is the length of z_{perm} and $z_{\text{perm},k}$ is the kth element in z_{perm}. Thus, the subtype with large P value (24) will not be considered as an interesting cancer subtype even though it has high $N01$.

To find clinical implication of the subtype, we evaluate the prognostic signature in the subgroup of patients using the logistic regression with L1 penalty. We first classify patients belonging to the subgroup into good and poor prognosis groups based on cross validated probabilities of being relapsed patients from the logistic regression. Then, the strength of the prognostic signatures from the logistic regression is assessed by computing the survival difference between good and poor prognosis groups and the area under the operating characteristic curve (AUC).

7. Real Data Analysis

7.1. Chemores Data Example. Lung cancer is one of the most prevalent and deadliest cancers. Human lung cancers are classified into two major subtypes, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC, which accounts for around 80% of all primary lung cancers, is a known heterogeneous group and its prognosis is generally poor [16]. In the current clinical practice, it is difficult to perform histopathological classification with small biopsies [17]. In order to improve the selection of patients who most likely will benefit from adjuvant chemotherapy (ACT), there is an urgent need to establish new diagnostic tools.

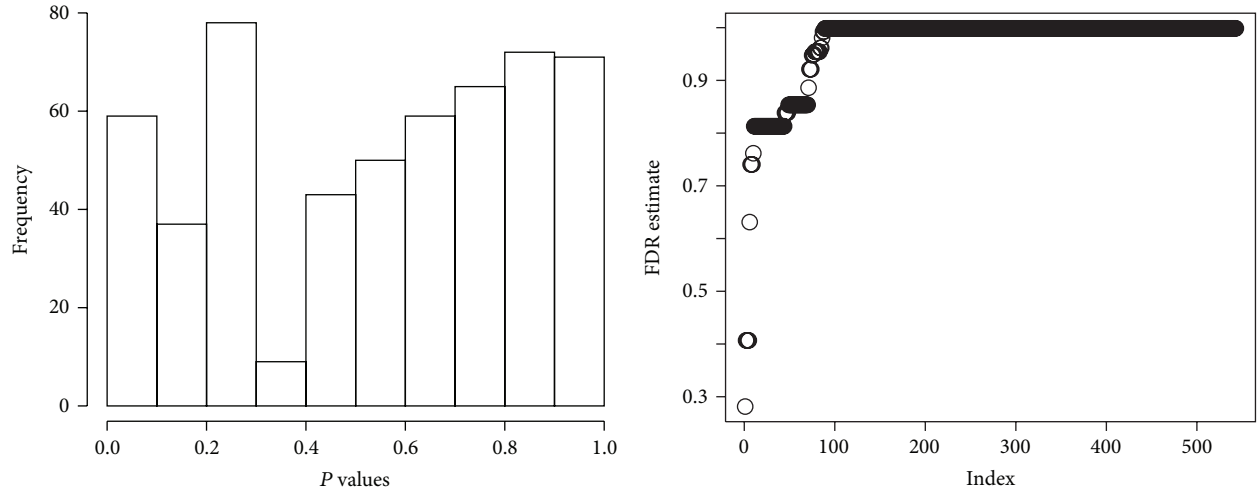
In this view, a study was organized by the Chemores initiative, which became an EU funded (FP6) Integrated Project involving 19 academic centers, organizations for cancer research, and research-oriented biotechnology companies in 8 European countries. Tissue samples from a cohort of 123 patients who underwent complete surgical resection between 30 January 2002 and 26 June 2006 are analyzed. All the patients belong to NSCLC and 59 patients experienced a relapse. This group of patients represents a heterogeneous group of lung cancers. We assayed the samples for gene expression, performed using dual-color human array from Agilent containing 41000 gene probes; a dye-swap of tumor versus normal lung tissue from same individual was employed for each sample and the log-ratio values were combined by averaging (the dataset is available at http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1132). Figure 5(a) shows the depletion in small P values of the two-sample t-statistics for the 41000 gene probes. Figure 5(b) shows the corresponding pessimistic standard FDR estimate by [11] (dashed line). Thus, to take into account the heterogeneity issue properly, CAMS is needed here. Two inputs for implementing CAMS are a gene expression matrix and a relapse indicator. Table 1 shows a summary of output. The first column of this output contains unique names of subtypes. The second and third columns tell how many genes are involved in defining each subtype and the number of patients in the subtype. The P values in the last column are computed using (24). The full lists of genes and patients can be identified by SubtypeID.

To reduce the computation time further, we consider filtering out uninteresting cases in the first stage. We compute $N01$ through the FDR based on the normal approximation only. We call this $N01_0$. If $N01_0$ is small, we skip the remaining procedure and go to search for next subtype. Figure 6(a) shows histogram of P values for $N01$ after filtering out the uninteresting cases having $N01_0 \le 2$. The standard FDR estimate is given in Figure 6(b), showing some interesting

Table 1: The output from *R* package.

Subtype_ID	Genes_in_clusters	Patients_in_subtype	*N*01	*P* value*
⋮	⋮	⋮	⋮	⋮
6	1535	69	17	0.020
7	124	28	23	0.020
⋮	⋮	⋮	⋮	⋮

*The *P* value of *N*01 for *i*th subtype is computed by using (24).



(a) Histogram of *P* values (b) FDR curve

Figure 6: (a) Histogram of *P* values for *N*01 from a lung cancer data and (b) the corresponding false discovery rate (FDR) estimate.

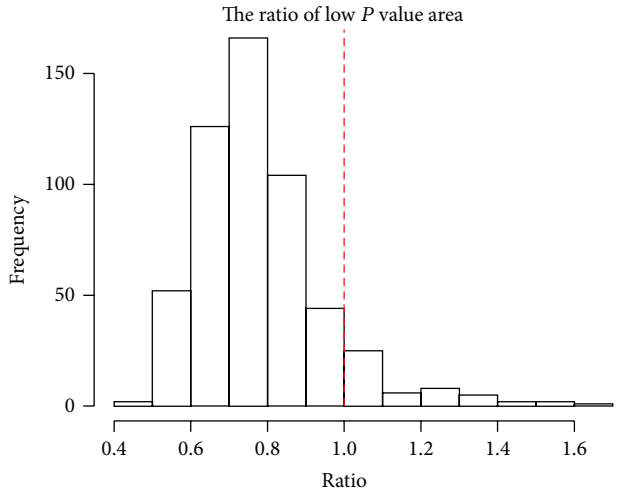
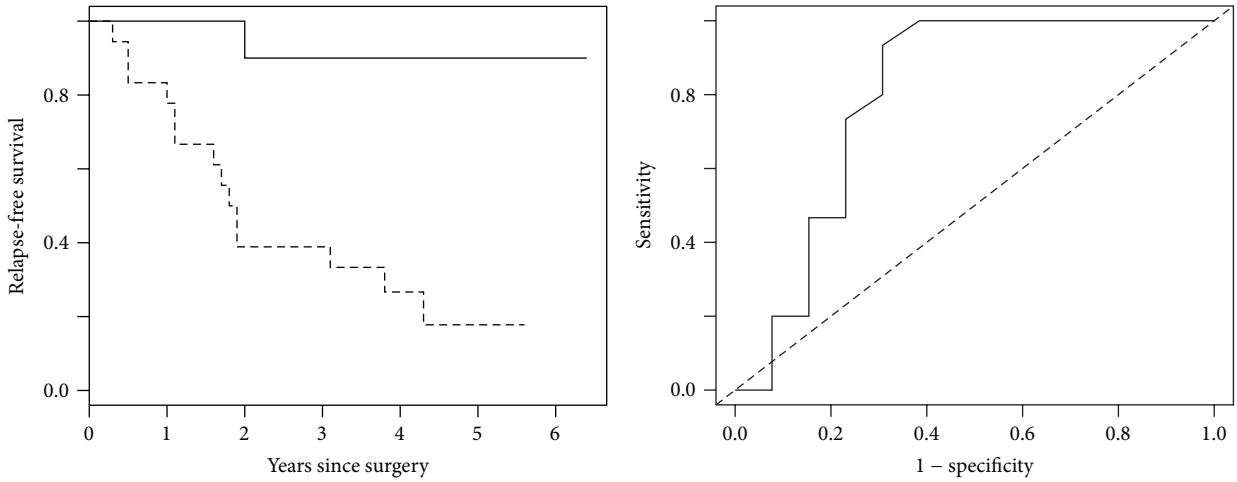


Figure 7: The proportion of subtypes showing depleted *P* value distribution from our clustering results is 0.908 (left side of vertical dashed line).

subtypes. In this analysis, we compute the proportion of subtypes showing depleted distributions with (22) and it is 0.908 (Figure 7). Therefore, *N*01 is essential in assessing the quality of each subtype.

From the top list of subtypes, one promising subtype is further analyzed using survival information to compute the appropriate prognostic signature for that subtype. To deal with large number of predictors (genes) we use logistic regression with L1 penalty [18] where the relapse status is the response variable. The cross validated probability of being a relapsed patient is computed from the leave-one-out cross validation, and the poor prognosis group is defined as the patients having the probability ≥0.5. To assess the strength of the prognostic signatures from the logistic regression, we compute the survival difference between good and poor prognosis groups. In Figure 8(a), the Kaplan-Meier curves of relapse-free survival show big difference between those two groups. Figure 8(b) shows operating characteristic curves for identifying relapse during follow-up. The area under the curve (AUC), computed under leave-one-out cross validation, is 0.806.

7.2. Bild et al.'s Data Example. As another application, we use lung cancer data by Bild et al. [19]. Their research purpose was to identify gene expression signatures of human cancers that reflect the activity of a given pathway. The gene expression dataset for lung cancer consists of 53 squamous cell carcinomas (SCC) and 58 adenocarcinomas (AC), so we expect that the group of patients represents a heterogeneous group. Among 58 relapsed patients, 26 and 32 patients belong to SCC and AC, respectively. The expression



(a) Kaplan-Meier Curve (b) ROC curve

Figure 8: (a) Kaplan-Meier curves of good and poor prognosis groups for a promising subtype and (b) receiver operating characteristic (ROC) curve.

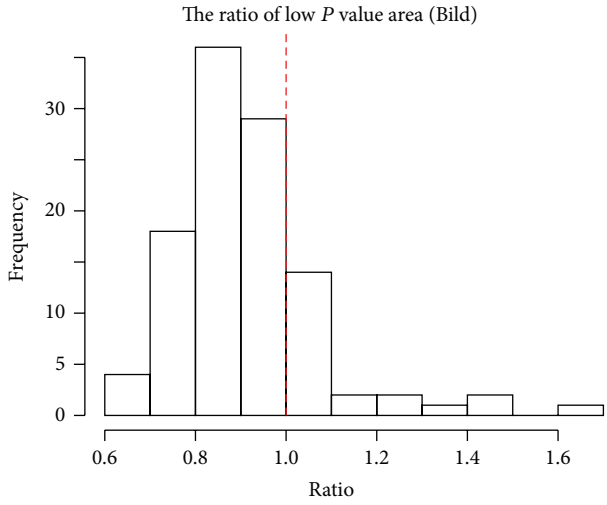


Figure 9: The proportion of subtypes showing depleted P value distribution from our clustering results is 0.798 (left side of vertical dashed line).

dataset was obtained using Human U133 2.0 plus arrays (Affymetrix) containing 56475 gene probes. It is available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3141. For the downstream analysis, we normalized the dataset for each patient to have zero mean after taking logarithm. The same procedures as described in analyzing Chemores data were applied to the normalized data. The proportion of subtypes showing depleted distributions is 0.798, so $N01$ is crucial in assessing the quality of each subtype. See Figure 9. Likewise in the previous section, further survival analysis can be done, but we omit the results here for brevity.

8. Discussion and Conclusions

In this paper, we proposed an extensive clustering algorithm to find cancer subtypes and have addressed the heterogeneity issue induced by the unobserved group to assess the resulting subtypes appropriately. The unobserved group creates a serious conservative bias problem when standard FDR estimation is used, but our proposed FDR estimation method resolves it. SVD is used as a tool for discovering the effect of heterogeneity on the null distribution of the test statistics. In particular, when many datasets are considered simultaneously, we develop a much faster and more stable FDR estimation algorithm than the method in [9].

Although we focus only on the heterogeneity issue in this paper, Efron's three issues [7] should be considered simultaneously in high-throughput data analysis. It is difficult, however, to distinguish genes with small effects from correlation effects because both can produce similarly wide distributions of the test statistic. We also expect that there is some confounding between the heterogeneity effect and the above two effects. Thus, careful joint approaches for dealing with the three issues are required. For example, Pawitan et al. [8] showed that it is possible to get less bias by estimating π_0 and $f_1(z)$ using a joint estimation method. This issue needs further investigation.

Recently, several biclustering algorithms have been proposed for gene expression data, and a comparative study was performed in [20]. They pointed out that performance on synthetic datasets did not always correlate with that on real datasets and no algorithm is uniformly the best under different environments. Considering this point, CAMS is also expected to have its own weakness and strength. Thus, it is needed to study when CAMS performs well compared to other biclustering methods. On the one hand, it is possible to embed existing biclustering algorithms into CAMS with some modification. Then, we can compare performances of various biclustering methods when subtypes are assessed by $N01$.

In addition to the above issues, there are still many scientific questions to be considered here. For example, should two similarly constructed clusters be combined or

remained separate? How can we assign an independent test sample to newly constructed subtypes? A practical method for dealing with these scientific problems will require further research.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

Yudi Pawitan, Woojoo Lee, and Andrey Alexeyenko conceived the study and wrote the paper. Woojoo Lee and Andrey Alexeyenko performed data analysis. All authors read and approved the final paper.

Acknowledgments

This work is supported by research grants from the European Union under the Chemores project and the Swedish Research Council and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1061332). The support by BILS (Bioinformatics Infrastructure for Life Sciences) is also gratefully acknowledged.

References

[1] C. M. Perou, T. Sørile, M. B. Eisen et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[2] T. Sørlie, C. M. Perou, R. Tibshirani et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10869–10874, 2001.

[3] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00)*, pp. 93–103, San Diego, Calif, USA, August 2000.

[4] A. Tanay, R. Sharan, and R. Shamir, "Biclustering algorithms: a survey," in *Handbook of Bioinformatics*, A. Srinivas, Ed., Chapman & Hall, New York, NY, USA, 2004.

[5] G. Nowak and R. Tibshirani, "Complementary hierarchical clustering," *Biostatistics*, vol. 9, no. 3, pp. 467–483, 2008.

[6] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, article e161, 2007.

[7] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.

[8] Y. Pawitan, K. R. K. Murthy, S. Michiels, and A. Ploner, "Bias in the estimation of false discovery rate in microarray studies," *Bioinformatics*, vol. 21, no. 20, pp. 3865–3872, 2005.

[9] Y. Pawitan, S. Calza, and A. Ploner, "Estimation of false discovery proportion under general dependence," *Bioinformatics*, vol. 22, no. 24, pp. 3025–3031, 2006.

[10] O. Stegle, L. Parts, R. Durbin, and J. Winn, "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies," *PLoS Computational Biology*, vol. 6, no. 5, 2010.

[11] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[12] C. Friguet, M. Kloareg, and D. Causeur, "A factor model approach to multiple testing under dependence," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1406–1415, 2009.

[13] J. T. Leek and J. D. Storey, "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 48, pp. 18718–18723, 2008.

[14] N. Meinshausen, "False discovery control for multiple tests of association under general dependence," *Scandinavian Journal of Statistics. Theory and Applications*, vol. 33, no. 2, pp. 227–237, 2006.

[15] B. Efron, "Microarrays, empirical bayes and the two-groups model," *Statistical Science*, vol. 23, no. 1, pp. 1–22, 2008.

[16] M. Tsuboi, T. Ohira, H. Saji et al., "The present status of postoperative adjuvant chemotherapy for completely resected non-small cell lung cancer," *Annals of Thoracic and Cardiovascular Surgery*, vol. 13, no. 2, pp. 73–77, 2007.

[17] P. T. Cagle, T. C. Allen, S. Dacie et al., "Revolution in lung cancer new challenges for the surgical pathologist," *Archives of Pathology and Laboratory Medicine*, vol. 135, no. 1, pp. 110–116, 2011.

[18] J. J. Goeman, "L_1 penalized estimation in the Cox proportional hazards model," *Biometrical Journal*, vol. 52, pp. 70–84, 2010.

[19] A. H. Bild, G. Yao, J. T. Chang et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.

[20] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V. Çatalyürek, "A comparative analysis of biclustering algorithms for gene expression data," *Briefings in Bioinformatics*, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 523641, 10 pages
http://dx.doi.org/10.1155/2015/523641

Research Article

Detecting Genetic Interactions for Quantitative Traits Using *m*-Spacing Entropy Measure

Jaeyong Yee,[1] Min-Seok Kwon,[2] Seohoon Jin,[3] Taesung Park,[4] and Mira Park[5]

[1]*Department of Physiology and Biophysics, Eulji University, Daejeon, Republic of Korea*
[2]*Department of Bioinformatics, Seoul National University, Seoul, Republic of Korea*
[3]*Department of Informational Statistics, Korea University, Jochiwon, Republic of Korea*
[4]*Department of Statistics, Seoul National University, Seoul, Republic of Korea*
[5]*Department of Preventive Medicine, Eulji University, Daejeon, Republic of Korea*

Correspondence should be addressed to Mira Park; mira@eulji.ac.kr

Received 14 November 2014; Revised 4 February 2015; Accepted 8 March 2015

Academic Editor: Xiang-Yang Lou

A number of statistical methods for detecting gene-gene interactions have been developed in genetic association studies with binary traits. However, many phenotype measures are intrinsically quantitative and categorizing continuous traits may not always be straightforward and meaningful. Association of gene-gene interactions with an observed distribution of such phenotypes needs to be investigated directly without categorization. Information gain based on entropy measure has previously been successful in identifying genetic associations with binary traits. We extend the usefulness of this information gain by proposing a nonparametric evaluation method of conditional entropy of a quantitative phenotype associated with a given genotype. Hence, the information gain can be obtained for any phenotype distribution. Because any functional form, such as Gaussian, is not assumed for the entire distribution of a trait or a given genotype, this method is expected to be robust enough to be applied to any phenotypic association data. Here, we show its use to successfully identify the main effect, as well as the genetic interactions, associated with a quantitative trait.

1. Introduction

Recent advances in high-throughput genotyping techniques have produced massive volumes of genetic data. Although it is common to analyze single SNP effects extensively, such approaches cannot adequately explain the intricate genetic contributions to complex diseases such as hypertension, diabetes, and certain psychiatric disorders. Consequently there are still large amounts of genetic components that remain unexplained. Gene-gene interaction analysis may be one method to adequately address this missing heritability problem [1].

For case-control studies, which formulate the measures for a binary trait, a number of statistical methods for detecting gene-gene interactions have been proposed. One of the most popular methods is multifactor dimensionality reduction (MDR) [2] that converts a high-dimensional contingency table to a one-dimensional model without raising the issue of sparse cells. Several variants of MDR have been recently developed [3–8], while another approach was developed [9–11] from information theory [12, 13]. More recently, an entropy-based approach which utilizes the relative gain of information, as well as its standardized measure, has also been proposed [14].

However, for quantitative traits such as the blood pressure, body mass index, and patient survival times, relatively few attempts have been made to analyze the genetic interactions. Because many phenotype measures are intrinsically quantitative, and categorizing a continuous trait may not always be straightforward and meaningful, association of gene-gene interactions with an observed distribution of such phenotypes needs to be investigated directly without categorization. To that end, introducing a new statistic is one way to tackle the problem [15]. Extending the MDR algorithm to continuous traits, as in the ways of the generalized MDR (GMDR) and the model-based MDR (MB-MDR), has been

proposed [3, 6]. More recently a quantitative MDR (QMDR) was proposed to replace the balanced accuracy metric with a t-test statistic [16]. However, these MDR-based approaches may oversimplify the original data to some degree, through classification of phenotypes. An entropy-based approach may well be an alternative model. Entropy is commonly used in information theory to measure the uncertainty of random variables [12, 13], and information gain or mutual information has been shown useful to represent association strengths [17–19]. Although the usefulness of such information theoretical methods is well known, the statistical methods based on this approach for analyzing gene-gene interactions of the quantitative traits are rarely found, with the exception of one specific case [20]. However, the application may also be limited by assuming a normal distribution.

Here, we extend the usefulness of the information concept to quantitative traits by considering nonparametric estimates based on sample-spacing or m-spacing [22–25] for the conditional entropy of a quantitative phenotype, based on a given genotype. The challenge, therefore, is to couple a nonparametric entropy estimator to correct and stable information gains. We thus developed the useful information gain standardized (IGS) approach and applied it to datasets composed of several genotypes and the quantitative trait. This approach could be considered an extension of previous work on categorical traits [14] to the quantitative phenotypes. The proposed method, however, does not attempt in any way to classify quantitative phenotypes like other methods, such as variants of MDR but instead handles them directly, providing an intrinsic advantage of removing the chance of misclassification. While previous entropy-based methods of analyzing quantitative traits assumed the shape of its distribution to be normal [20], our method does not need to specify the distribution to estimate the association. Any regular or irregular distribution would not cause any difficulties. Although this is also an advantage of GMDR or QMDR, we propose a method that takes the advantageous characteristics from both of those methods. We also performed extensive simulation studies to compare the powers of the proposed method to QMDR and GMDR, demonstrating its advantage in detection power.

In the following sections, after a brief review of nonparametric entropy estimation, we describe a new method for modeling genetic interactions. A nonparametric entropy estimator is shown to successfully couple with genetic datasets through our modifying work in the Materials and Methods. Application of this information gain standardized (IGS) approach is evaluated for both simulation and real datasets in the Results and Discussions.

2. Materials and Methods

2.1. Estimation of the Entropy for a Continuous Variable. If X is a random vector with probability density function, $f(x)$, its differential entropy is defined by

$$H(f) = -\int f(x) \ln(f(x))\, dx. \tag{1}$$

A well-known approach for estimating a solution to this equation is to use plug-in estimates. In this approach, $f(x)$ is first estimated using a standard density estimation method such as a histogram or kernel density estimator, and the entropy is then computed. Integral, resubstitution, splitting data, and cross-validation estimates are among the usual plug-in estimates [22]. Another approach is based on sample-spacing. Let $\{X_k\}$ be a set of independent and identically distributed real valued random variables, with corresponding order statistics of $\{X_{n,k}\}$. Here, n represents the total number of measured samples. For the arbitrary integers i and m satisfying the condition of $1 \le i < i + m \le n$, a spacing of order m or m-spacing is defined as $X_{n,i+m} - X_{n,i}$. A density estimate, based on sample-spacing, m, is then constructed as

$$f_n(x) = \frac{m}{n} \frac{1}{X_{n,im} - X_{n,(i-1)m}}, \tag{2}$$

where $x \in [X_{n,(i-1)m}, X_{n,im})$ [14]. This density estimate is consistent if, as $n \to \infty$, $m \to \infty$ and $m/n \to 0$ [22]. Several variations of an entropy estimator with minor differences have been proposed, all based on the above density estimates [23, 24]. Among them, the following were reported to approximate with lowered variance [25]:

$$H_{m,n} = \frac{1}{n-m} \sum_{k=1}^{n-m} \ln\left(\frac{n}{m}\left(X_{n,k+m} - X_{n,k}\right)\right). \tag{3}$$

Asymptotic bias of this estimator can be corrected by adding additional terms, including the digamma function [22, 28]:

$$H_{m,n} = \frac{1}{n-m} \sum_{k=1}^{n-m} \ln\left(\frac{n}{m}\left(X_{n,k+m} - X_{n,k}\right)\right) - \frac{\Gamma'(m)}{\Gamma(m)} + \ln m. \tag{4}$$

As m increases, the correctional terms become negligible and the two estimators coincide. Our evaluation of the entropy of a phenotype, $H(P)$, of a quantitative trait is based on this estimator.

2.2. Modification of the m-Spacing Based Entropy Estimator. The estimator in (4) has both n and m as parameters. In genetic association studies, the number of samples, n, of several hundreds is common. However, when the conditional entropy is estimated, there may be a minor allele that could have a much smaller number of samples corresponding to that allele. Moreover, the choice of the sample-spacing, m, should affect the resulting estimation of an entropy value. Therefore, it is required to have an entropy estimation scheme independent of the number of samples, without the need of choosing a particular value of the sample-spacing. To illustrate such a requirement, an ensemble of 3,000 sets of the random deviation from $N(0, 1^2)$ was generated for each data point in Figure 1, where the mean and standard deviation of the estimates are plotted for each ensemble. On the left panel of Figure 1, m is fixed to 10 and 20 while n is varied. The analytic formula of the entropy for a normal distribution can be obtained as follows [20], where e is Euler's number:

$$H = \ln\left(\sigma\sqrt{2\pi e}\right). \tag{5}$$

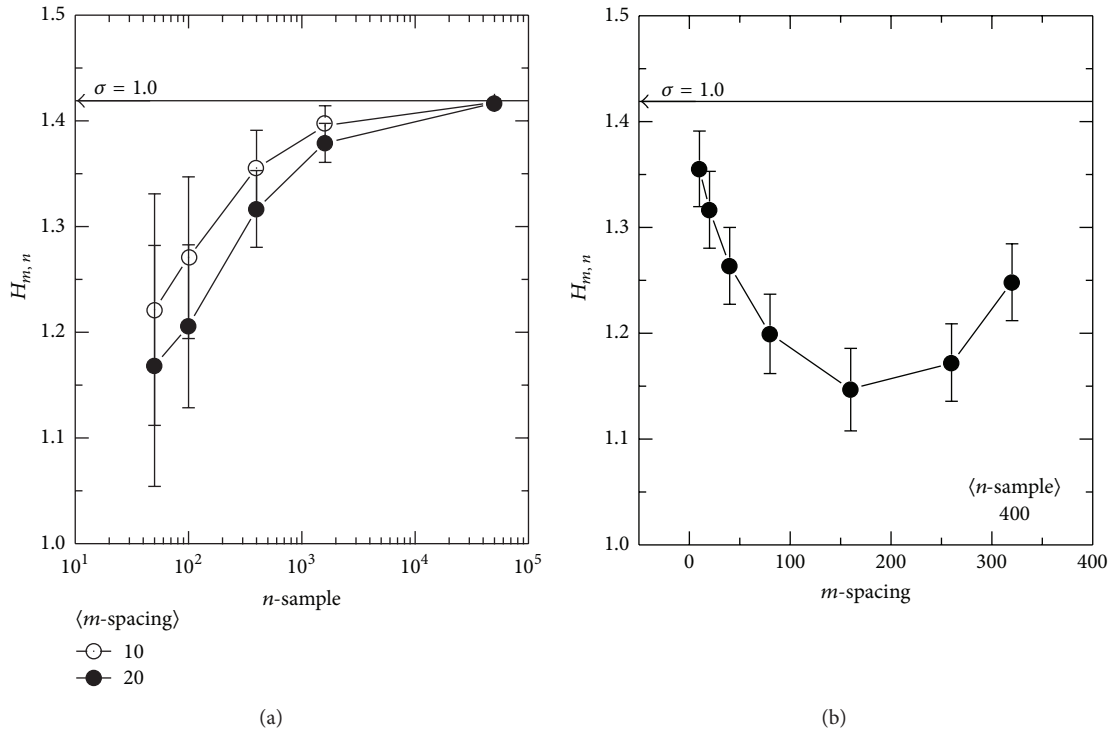


Figure 1: The n-dependence (a) and m-dependence (b) of the entropy estimator $H_{m,n}$. An ensemble of 3,000 sets of random sampling from $N(0, 1^2)$ was constructed and used for each point in the plot. The sample-spacing, m, was fixed while varying the number of samples, n, (a) to evaluate the n-dependence of the entropy estimator. In (b), n was fixed and m was varied to show the m-dependence. Analytically obtained true values are represented by the arrowed horizontal lines.

The calculated value of (5) is pointed on the vertical axis with a horizontal arrow with the corresponding σ above it. The obvious n-dependence of the estimator can be seen in this plot, where the estimation approaches the analytic value, as n increases with $\sqrt{n}$-consistency, as expected [24]. In Figure 1(b), n is fixed to 400, while m is varied. In this plot, the estimated entropy again changes in value throughout the possible range of m. It is shown that the estimated value is always smaller than the analytically calculated value. Therefore, assigning a particular value to m such as $\sqrt{n}$, the typical choice [25], would not be appropriate in this sampling range. Because of these n- and m-dependences, the estimator in (4) may need to be modified. Therefore, we modify the entropy estimator in (4) as follows:

$$H_{\langle m\rangle,n} = \frac{1}{n-1}\sum_{m=1}^{n-1}\left(\frac{1}{n-m}\sum_{k=1}^{n-m}\ln\left(\frac{n}{m}\left(X_{n,k+m}-X_{n,k}\right)\right) - \frac{\Gamma'(m)}{\Gamma(m)} + \ln m\right). \tag{6}$$

In this modification, an entropy estimator is averaged over the possible m values for each n, which is denoted by $\langle m\rangle$. This estimator is used to plot the entropy versus number of samples in Figure 2. Over a wide range of n, this entropy estimator yields very stable values, in contrast to Figure 1(a). An increase in the extremely small n range should be within the tolerable error in an application of genome-wide association, as the contribution to the conditional entropy by such a minor allele would be suppressed by the weighting factor of the marginal probability that should be proportional to the number of corresponding samples. Analytically obtained entropy values for $N(0, \sigma^2)$, with three different σ's, are marked on the vertical axis on the right-hand side. Regardless of the value of σ, the differences between the analytically obtained value and the values given by the estimator stay essentially the same. Considering that the association study measures the difference between the entropy and the corresponding conditional entropy, the stability should be a more critical issue than the absolute value of the estimates. Therefore compensation of this Δ would not be necessary as long as it is stable. Furthermore, the underestimation of the entropy shown in the plot should have little effect on the association strength. Hence, an entropy estimator has been set up that should satisfy the practical n-independence without the need to find a proper sample-spacing.

2.3. Evaluation of a Conditional Entropy. Now let G be a categorical variable assigned to each sample measurement X_k. G may be a genotype given by a measured SNP or a combination of SNPs, while X_k represents the measured value of a phenotype. For detecting the main effect of a single SNP, G consists of three categories of $G = 0$, $G = 1$, and $G = 2$. For detecting the interaction between SNP_i and SNP_j, G consists of 9 categories, such that $G = 0 = (\mathrm{SNP}_i = 0, \mathrm{SNP}_j = 0)$,

$G = 1 = (\mathrm{SNP}_i = 0, \mathrm{SNP}_j = 1)$, $G = 2 = (\mathrm{SNP}_i = 0, \mathrm{SNP}_j = 2)$, $G = 3 = (\mathrm{SNP}_i = 1, \mathrm{SNP}_j = 0), \ldots$, and $G = 8 = (\mathrm{SNP}_i = 2, \mathrm{SNP}_j = 2)$. Detection of the higher order interaction can be performed in the same way with expansion of the categories of G. Then an estimator for each specific component of the conditional entropy, $H(P \mid G = g)$, can be constructed using the genotype-selected subset measurements $\{X_{n_g,k}\}$, along with an individual sample-spacing of m_g. Extending (6), while applying the above argument, should now readily produce the estimators for the entropy of a phenotype and the conditional entropy. Here d denotes the order of a gene-gene interaction:

$$H(P) = \frac{1}{n-1}\sum_{m=1}^{n-1}\left(\frac{1}{n-m}\sum_{k=1}^{n-m}\ln\left(\frac{n}{m}\left(X_{n,k+m} - X_{n,k}\right)\right) - \frac{\Gamma'(m)}{\Gamma(m)} + \ln m\right),$$

$$H(P \mid G) = \sum_{g=0}^{3^d-1}\frac{n_g}{n}\left(\frac{1}{n_g-1}\sum_{m_g=1}^{n_g-1}\left(\frac{1}{n_g-m_g}\cdot\sum_{k=1}^{n_g-m_g}\ln\left(\frac{n_g}{m_g}\left(X_{n_g,k+m_g} - X_{n_g,k}\right)\right) - \frac{\Gamma'\left(m_g\right)}{\Gamma\left(m_g\right)} + \ln m_g\right)\right) = \sum_{g=0}^{3^d-1}\frac{n_g}{n}H\left(P \mid G = g\right). \tag{7}$$

2.4. Standardized Measure of an Association Strength. Since the differential entropy values are scale-dependent, when the above estimators are calculated with $\{X_i\}$ and $\{cX_i\}$ (where c is a constant scale factor), the difference would be $\ln c$:

$$H_{\{cX_i\}} = \ln c + H_{\{X_i\}}. \tag{8}$$

For example, if the phenotype is height it may be measured in meters or centimeters. In this case, the scale factor is 100. Nevertheless, the association strength should also be the same. Also note that a negative value is perfectly legitimate for a differential entropy. Information gain, IG, as in the form defined with discrete entropies [14], should satisfy scale independence, while correctly representing an association strength without being affected by negative values. Therefore, it should retain its usefulness as a measure of an association strength:

$$\mathrm{IG} = H(P) - H(P \mid G). \tag{9}$$

IG would be readily estimated with the above estimator (7). IG standardized (IGS) is set up with the means and standard deviations of IGs obtained from repeated shuffling of the phenotypes while all genotypes remained fixed [14].

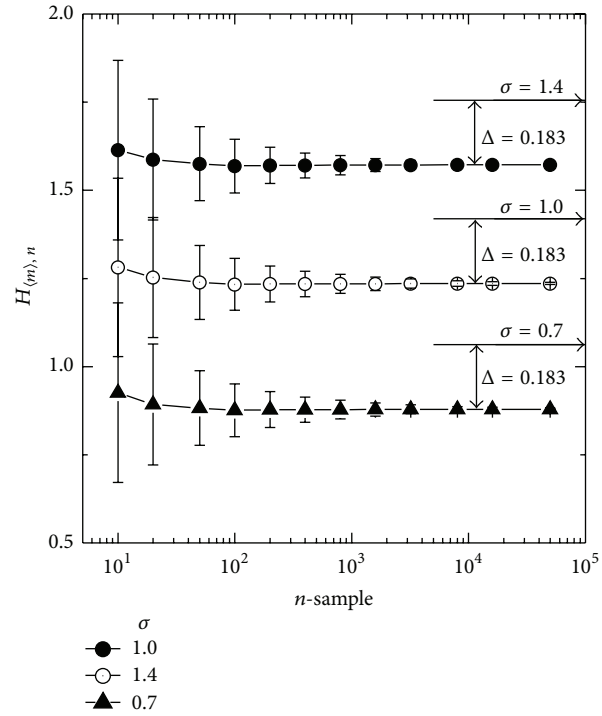


Figure 2: The n-independence and constant offset from the true value of the estimates averaged over all possible m values for each n. Each symbol represents a result of samplings from $N(0, \sigma^2)$. While varying n, the number of samples, the estimated entropy values were averaged over all the possible m, sample-spacing values. $\langle m\rangle$ denotes this averaging, which should not depend on weighting due to the virtually same standard deviations shown in Figure 1(b). Over a wide range of n, the estimated entropy stays effectively the same, showing n-independence in the range of practical number of sampling. Moreover, the almost flat line connecting each symbol shifts up or down following exactly the change of the true value indicated by the horizontal arrows. The rise in the extremely small n range should be within the tolerable error of any specific application, because the contribution to conditional entropy by such a case would be suppressed by weighting, based on the marginal probability that should be proportional to n.

Let $\mathrm{IG}_i^{(1)}$ denote the maximum IG of the ith permuted dataset. Then, the mean and standard deviation of $\mathrm{IG}_1^{(1)}, \mathrm{IG}_2^{(1)}, \ldots, \mathrm{IG}_n^{(1)}$ can be computed as follows:

$$\overline{\mathrm{IG}}_p = \frac{\sum_{i=1}^{n}\mathrm{IG}_i^{(1)}}{n}, \qquad S_p = \sqrt{\frac{\sum_{i=1}^{n}\left(\overline{\mathrm{IG}}_i^{(1)} - \overline{\mathrm{IG}}_p\right)^2}{n-1}}, \tag{10}$$

where n is the number of permuted datasets. Now IGS is defined as follows:

$$\mathrm{IGS} = \frac{\mathrm{IG} - \overline{\mathrm{IG}}_p}{S_p}. \tag{11}$$

3. Results and Discussions

3.1. Demonstration of the m-Spacing Method. To show the plausibility of the proposed m-spacing method,

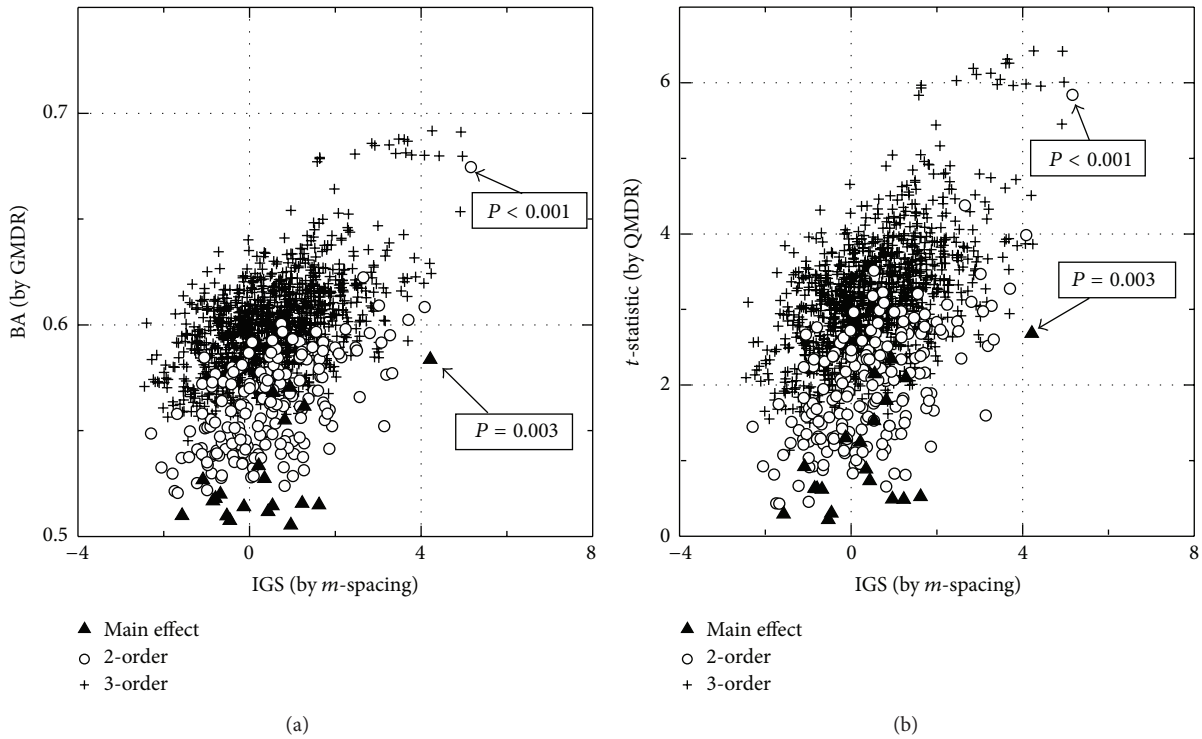


Figure 3: Comparison of the QMDR, GMDR, and m-spacing methods. Association strengths obtained by GMDR versus m-spacing (a) and by QMDR versus m-spacing (b) are compared for a simulated dataset. All three methods were used to evaluate the main effect as well as 2nd and 3rd order interactions. The dataset was designed to have one 2nd order interaction causal pair.

a representative result is shown in Figure 3, using a dataset whose quantitative trait was generated from a normal distribution with a single causal SNP pair simulated, as described in the next section. The sample size of the dataset was 400, with 20 SNPs. In panel (a), the association strengths, obtained by m-spacing and GMDR, are plotted as horizontal and vertical coordinates, respectively. Filled triangles represent the main effects, while open circles are for the 2nd order interactions. Both methods identify the same single SNP pair having a prominent interaction plotted in the upper right corner. One of the SNPs was found to produce the main effect, in contrast to others. Again, the result is agreed by both methods. P values obtained by permutation are given in the boxes for those selected points. Association strengths of the 3rd order interactions are plotted with a plus sign. Because no 3rd order interaction is simulated into the dataset, the combinations of SNPs made by adding a single SNP to the causal pair are expected to have high association values. Those points are clustered near the identified causal pair in the upper right corner. In panel (b) of Figure 3, the same comparison was made using the result from m-spacing and QMDR. Both comparisons show consistent results between the proposed m-spacing method and GMDR or QMDR. Note that IGS instead of IG was used. The distribution of the IG values from a dataset would shift to a higher direction, with increased order of interactions. Thus, the more conditions applied, the less entropy may be left to find. In other words, as the order of interaction increases, the conditional entropy $H(P \mid G)$ tends to decrease, while $H(P)$ remains the same. Therefore IGS is vital if one needs to compare the association strengths between genotypes from different orders of interactions. Figure 3 shows that the simulated causal pair has the largest IGS value among all points, from different orders of interactions.

3.2. Generation of the Simulation Data. To examine the performance of the m-spacing method, an extensive set of simulation data was necessarily generated. First, three types of quantitative trait distributions were considered. Two of them were normal and gamma distributions, and another one was a mixture of those two types. With single causal pair designed, 70 different penetrance models, based on [21], were incorporated. For the case of a normal distribution, a phenotype value, y, associated with two interacting SNPs was selected from a normal distribution, as defined by the penetrance values tabled for possible combinations of genotypes associated as follows:

$$y \mid (\mathrm{SNP}_1 = i, \mathrm{SNP}_2 = j) \sim N\left(f_{ij}, \sigma^2\right). \tag{12}$$

Here f_{ij} represents the penetrance values tabled for every model simulated and can be found in [21]. It is tabulated for each possible pair of genotypes, (i, j). In 70 different penetrance models, 14 different combinations of two different minor allele frequencies (MAFs) and seven different heritability values were considered. Specifically, we considered the cases when the MAFs were 0.2 and 0.4 and when the heritability was 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, and 0.4. Three

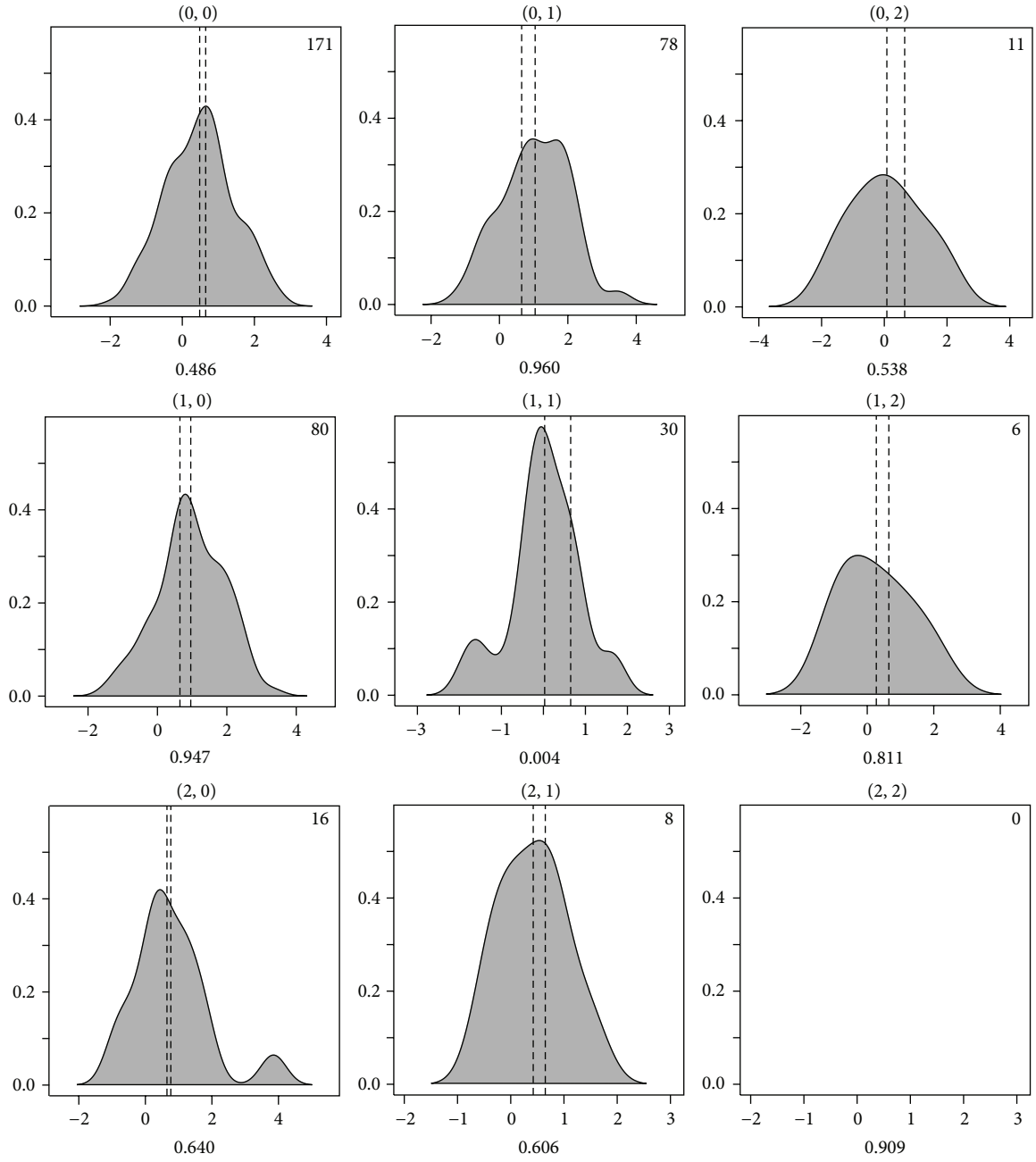


Figure 4: Demonstration of the simulation scheme. Phenotype distributions were plotted to associate with the genotypes by two interacting SNPs, as denoted in the parentheses on top of each plot. SNPs may take values of 0, 1, and 2 or AA, Aa, and aa. For this particular dataset, the MAF was set to 0.200. On the bottom of each plot, the penetrance value for this particular model is given, which is taken from [21]. Inside each plot, the number of samples generated to satisfy the simulation constraint is given. The vertical dotted lines are for the mean values of the high- and low-risk groups. By constraint, the line on the left is for the low-risk group.

different values (0.8, 1.0, and 1.2) of the variance, σ, were used independently for the high- and low-risk groups, resulting in 9 combinations. The grouping constraint for the generated event was set such that the averaged y of the high-risk group should be larger than or equal to the overall average. The averaged y of the low-risk group should be less than the overall average. In Figure 4, 9 possible distributions of a generated phenotype are shown. In this example, the sample size is 400. The high- and low-risk groups have the same number of samples and both have a variance of 1.0. For gamma distributions, phenotype values follow the rule below:

$$y \mid \left(\mathrm{SNP}_1 = i, \mathrm{SNP}_2 = j\right) \sim \Gamma\left(k, \theta\right). \tag{13}$$

The shape and scale parameters, k and θ, were determined by f_{ij} and σ, using the relationship $f_{ij} = k\theta$ and $\sigma = k\theta^2$. Penetrance models were classified by 7 heritability values:

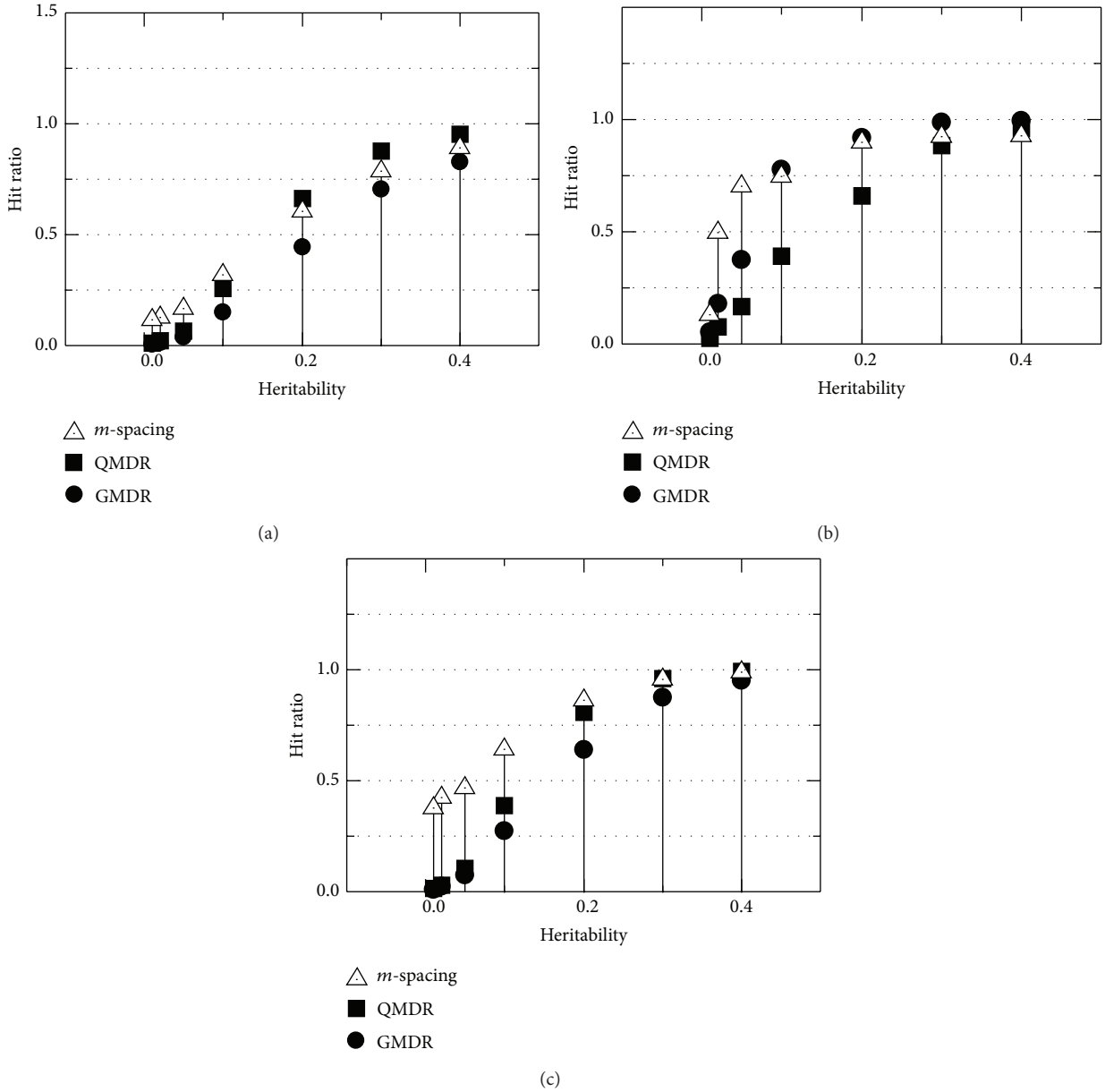


Figure 5: Comparison of the hit ratios or the detection probabilities among the proposed m-spacing method, QMDR, and GMDR. Genomic datasets were generated based on 70 different penetrance functions [21], which were, in turn, classified into 7 distinct values of heritability. For each model, the phenotype values are simulated with normal (a), gamma (b), and mixed (c) distributions. High- and low-risk groups in a quantitative trait overlapped with 9 different combinations of the standard deviations. Considering all of the above, 100 data files were generated for each case, adding up to 9,000 simulated files being examined for each point in the plot.

0.01, 0.02, 0.05, 0.1, 0.2, 0.3, and 0.4, resulting in 10 models for each heritability. The generated data files had a sample size of 400, with 20 SNPs. In all, $3 \times 70 \times 9 = 1{,}890$ different conditions were set up, with 100 simulated data files generated for each condition.

3.3. Comparison of the Detection Probability and Type I Error. The "hit ratio," or detection power, of the IGS was evaluated and compared. Simulated data files described in the previous subsection were used. All of them had a single causal pair to identify. In addition to our proposed m-spacing method, QMDR and GMDR were used to compare the results. Figure 5 shows the comparison. Panels (a), (b), and (c) are for the quantitative trait of normal, gamma, and mixed distributions, respectively. Seventy penetrance models were grouped into 7 cases of heritability on the horizontal axis, while all 9 combinations of the variances in high- and low-risk distributions were merged into each heritability case. With a normal distribution, as shown in Figure 5(a), the m-spacing's performance was in between those of QMDR and GMDR for higher values of penetrance. However, in the range of penetrance less than 0.2, the m-spacing performs

best. Note that the QMDR shows higher detection probability than the GMDR throughout the range. In the case of a gamma distribution, as shown in Figure 5(b), the QMDR's performance drops rapidly, as the heritability decreases when the hit ratios of m-spacing, as well as the GMDR, stay better than that of QMDR and are comparable to each other. Note the switch of the GMDR and QMDR's performance ranks with the change of the phenotype distribution. What QMDR does is essentially the dichotomization of the observed values of the quantitative phenotypes. Therefore, it should do better with well-defined symmetric distributions, such as a normal distribution, than with an asymmetric one (e.g., gamma distribution). The proposed m-spacing method is expected to be effective regardless of the shape of the phenotype distribution, because it makes no assumptions regarding the distribution and is therefore nonparametric, as demonstrated in Figures 5(a) and 5(b). This nonparameterization is again confirmed in Figure 5(c), showing that m-spacing outperforms the QMDR and the GMDR, throughout the whole range of heritability, in the case of the mixed form of phenotype distribution. Among the three methods examined, m-spacing was the most robust, performing consistently within the range of conditions for the simulation.

To estimate the type I error rate, the null datasets were generated under the same scheme as used for the detection power analysis except that there was no causal pair intended. Now there are 20 SNPs that none of the pairs are expected to have an association. Permutation P values for a particular pair were obtained by permuting each dataset 1000 times. We took the significance level α as 0.05 to get the ratio of the permutation P values smaller than or equal to α. We report this ratio as the type I error rate in Table 1, whose accuracy to one decimal place when expressed in percent was ensured by the number of the permutation. Table 1 presents the type I error rate for each combination of three trait distributions, two MAFs, and seven heritability values, along with the overall estimates. Throughout these conditions, the type I error rates are gathered tightly around 5% with maximum and minimum of 5.4% and 4.3%, respectively. Moreover there exists no sign of the dependence on the trait shape, heritability, and MAF. Therefore our proposed method preserved the type I error rates on these conditions.

3.4. Application to Real Data. A full-scale real dataset from the Korean Association Resource (KARE) project [20] was analyzed to investigate the effectiveness of the m-spacing method. Among the available phenotypes, "height" was chosen with a sample size of 8,842 from the population-based cohort. The total number of SNPs was 327,872, spanning over 22 chromosomes. The "height" phenotype showed to be close to a normal distribution such that the m-spacing method may not take advantage of the shape of the phenotype distribution, as discussed in the previous subsection. Table 2 lists the SNPs, selected by the m-spacing method (IGS), that had the strongest main effects. Out of 10 selected SNPs, rs2079795 and rs6440003 coincide with two previous reports [26, 27], although two more matched SNPs, rs11989122 and rs1344672, could be found as results of our analysis using the same tool as in [26], but using the newly imputed dataset. P values were estimated by permutation of the phenotype values to make null distributions. Permutations were iterated 100,000 and 10,000 times for the main effect and the interaction, respectively. A clear distinction between rs11989122 and the other selected SNPs can be seen in the IGS values. In Table 3, the 2nd order gene-gene interaction result is given. The top selected pair (rs6499786, rs1788421) was found to have the strongest association with "height," but the distinction was not so obvious, compared to the case of the main effect.

Table 1: Type I error estimation with the significance level α of 0.05.

Type I error rate (%)		Normal	Gamma	Mixed
MAF	0.2	5.0	5.0	5.1
	0.4	5.1	5.0	5.1
Heritability	0.01	5.3	5.0	4.8
	0.02	4.9	5.4	5.2
	0.05	5.3	4.3	5.3
	0.1	5.0	5.3	5.1
	0.2	5.0	5.3	5.1
	0.3	4.8	4.9	4.8
	0.4	5.1	4.7	5.3
Overall		5.0	5.0	5.1

4. Conclusion

In this paper, we present a modified m-spacing method for genome-wide association studies with a quantitative trait. The robustness of this method makes it useful for a wide range of sample sizes, while the original m-spacing method yields a reliable result only for datasets with a large sample size. Extensive simulation was performed to produce the datasets with different shapes of phenotype distributions, while varying the penetrance functions and adjusting the heritability as well. Causal pair detection probability was unaffected the most by the compared methods, based on the distribution shape and heritability, while GMDR and QMDR showed more dependency. The proposed m-spacing method is proven to outperform the others regardless of the shape of the trait distribution and also the range of lower heritability. In the higher heritability region, the performance of the proposed method is comparable to that of GMDR or QMDR, whichever shows better performance in that region. This would lead to versatile applicability of our nonparametric method for quantitative traits, with various characteristics. We applied this method to successfully identify the main effect and gene-gene interactions for the phenotype "height" with the full set of KARE samples. Although several of them overlapped with a previous report, new interactions were also found. Because "height" is presumed to be a trait with a normal distribution having a higher heritability, our method may be said to have performed successfully with no advantage over other methods. More extensive study is needed for quantitative traits, having various characteristics, to further demonstrate the expected robustness of our modified m-spacing method.

Table 2: Application of the m-spacing method to a full set of KARE samples with the phenotype "height;" main effect.

		Main effect		
rs ID	Chromosome	IGS	P value	Previous report
rs11989122	8	11.3892	1×10^{-5}	$^{*}5.89 \times 10^{-6}$
rs7316119	12	8.7531	1×10^{-5}	—
rs936634	18	8.6125	2×10^{-5}	—
rs7632381	3	7.8235	1×10^{-5}	—
rs2079795	17	7.6542	1×10^{-5}	2.92×10^{-6} Ref. [26]
rs1344672	3	7.6177	1×10^{-5}	$^{*}5.21 \times 10^{-7}$
rs2523865	6	7.6044	4×10^{-5}	—
rs3790199	20	7.5362	2×10^{-5}	—
rs6440003	3	7.5231	1×10^{-5}	3.87×10^{-7} Ref. [27]
rs17628655	19	7.5117	6×10^{-5}	—

*Identified using the same method as [26] but with imputed data, which is the same one we analyzed.

Table 3: Application of the m-spacing method to a full set of KARE samples with the phenotype "height;" 2nd order interaction.

		2nd order interaction			
rs ID	Chromosome	rs ID	Chromosome	IGS	P value
rs6499786	16	rs1788421	21	4.6197	1×10^{-4}
rs2529232	7	rs1788421	21	4.3869	1×10^{-4}
rs2241704	19	rs1788421	21	4.3855	1×10^{-4}

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (NRF-2013R1A1A2062848, NRF-2012R1A3A2026438).

References

[1] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 4, pp. 1193–1198, 2012.

[2] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[3] X.-Y. Lou, G.-B. Chen, L. Yan et al., "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence," *American Journal of Human Genetics*, vol. 80, no. 6, pp. 1125–1137, 2007.

[4] Y. Chung, S. Y. Lee, R. C. Elston, and T. Park, "Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions," *Bioinformatics*, vol. 23, no. 1, pp. 71–76, 2007.

[5] S. Yeoun Lee, Y. Chung, R. C. Elston, Y. Kim, and T. Park, "Log-linear model-based multifactor dimensionality reduction method to detect gene-gene interactions," *Bioinformatics*, vol. 23, no. 19, pp. 2589–2595, 2007.

[6] M. L. Calle, V. Urrea, G. Vellalta, N. Malats, and K. V. Steen, "Improving strategies for detecting genetic patterns of disease susceptibility in association studies," *Statistics in Medicine*, vol. 27, no. 30, pp. 6532–6546, 2008.

[7] W. S. Bush, T. L. Edwards, S. M. Dudek, B. A. McKinney, and M. D. Ritchie, "Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction," *BMC Bioinformatics*, vol. 9, article 238, 2008.

[8] K. Kim, M.-S. Kwon, S. Oh, and T. Park, "Identification of multiple gene-gene interactions for ordinal phenotypes," *BMC Medical Genomics*, vol. 6, supplement 2, article S9, 2013.

[9] G. Kang, W. Yue, J. Zhang, Y. Cui, Y. Zuo, and D. Zhang, "An entropy-based approach for testing genetic epistasis underlying complex diseases," *Journal of Theoretical Biology*, vol. 250, no. 2, pp. 362–374, 2008.

[10] C. Dong, X. Chu, Y. Wang et al., "Exploration of gene-gene interaction effects using entropy-based methods," *European Journal of Human Genetics*, vol. 16, no. 2, pp. 229–235, 2008.

[11] C. Wu, S. Li, and Y. Cui, "Genetic association studies: an information content perspective," *Current Genomics*, vol. 13, no. 7, pp. 566–573, 2012.

[12] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[13] R. M. Gray, *Entropy and Information Theory*, Springer, New York, NY, USA, 2nd edition, 2011.

[14] J. Yee, M.-S. Kwon, T. Park, and M. Park, "A modified entropy-based approach for identifying gene-gene interactions in case-control study," *PLoS ONE*, vol. 8, no. 7, Article ID e69321, 2013.

[15] M. Li, C. Ye, W. Fu, R. C. Elston, and Q. Lu, "Detecting genetic interactions for quantitative traits with *U*-statistics," *Genetic Epidemiology*, vol. 35, no. 6, pp. 457–468, 2011.

[16] J. Gui, J. H. Moore, S. M. Williams et al., "A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits," *PLoS ONE*, vol. 8, no. 6, Article ID e66545, 2013.

[17] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[18] N. N. Schraudolph, "Grandient-based manipulation of non-parametric entropy estimates," *IEEE Transactions on Neural Networks*, vol. 14, no. 2, pp. 1–10, 2004.

[19] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1415–1438, 2003.

[20] P. Chanda, L. Sucheston, S. Liu, A. Zhang, and M. Ramanathan, "Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits," *BMC Genomics*, vol. 10, article 1471, pp. 509–530, 2009.

[21] D. R. Velez, B. C. White, A. A. Motsinger et al., "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiology*, vol. 31, no. 4, pp. 306–315, 2007.

[22] J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. van der Meulen, "Nonparametric entropy estimation: an overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[23] A. B. Tsybakov and E. C. van der Meulen, "Root-n consistent estimators of entropy for densities with unbound support," *Scandinavian Journal of Statistics*, vol. 23, pp. 75–83, 1992.

[24] F. El Haje Hussein and Y. Golubev, "On entropy estimation by *m*-spacing method," *Jounal of Mathematical Sciences*, vol. 163, pp. 290–309, 2009.

[25] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *Journal of Machine Learning Research*, vol. 4, pp. 1271–1295, 2004.

[26] Y. S. Cho, M. J. Go, Y. J. Kim et al., "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits," *Nature Genetics*, vol. 41, no. 5, pp. 527–534, 2009.

[27] M. N. Weedon, H. Lango, C. M. Lindgren et al., "Genome-wide association analysis identifies 20 loci that influence adult height," *Nature Genetics*, vol. 40, no. 5, pp. 575–583, 2008.

[28] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *American Journal of Mathematical and Management Sciences*, vol. 23, no. 3-4, pp. 301–321, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 671859, 7 pages
http://dx.doi.org/10.1155/2015/671859

Research Article

A Comparative Study on Multifactor Dimensionality Reduction Methods for Detecting Gene-Gene Interactions with the Survival Phenotype

Seungyeoun Lee,[1] Yongkang Kim,[2] Min-Seok Kwon,[3] and Taesung Park[2,3]

[1]*Department of Mathematics and Statistics, Sejong University, Seoul 143-747, Republic of Korea*
[2]*Department of Statistics, Seoul National University, Seoul 151-747, Republic of Korea*
[3]*Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Republic of Korea*

Correspondence should be addressed to Seungyeoun Lee; leesy@sejong.ac.kr and Taesung Park; tspark@snu.ac.kr

Received 21 November 2014; Revised 18 April 2015; Accepted 27 April 2015

Academic Editor: Xiang-Yang Lou

Genome-wide association studies (GWAS) have extensively analyzed single SNP effects on a wide variety of common and complex diseases and found many genetic variants associated with diseases. However, there is still a large portion of the genetic variants left unexplained. This missing heritability problem might be due to the analytical strategy that limits analyses to only single SNPs. One of possible approaches to the missing heritability problem is to consider identifying multi-SNP effects or gene-gene interactions. The multifactor dimensionality reduction method has been widely used to detect gene-gene interactions based on the constructive induction by classifying high-dimensional genotype combinations into one-dimensional variable with two attributes of high risk and low risk for the case-control study. Many modifications of MDR have been proposed and also extended to the survival phenotype. In this study, we propose several extensions of MDR for the survival phenotype and compare the proposed extensions with earlier MDR through comprehensive simulation studies.

1. Introduction

In early genome-wide association studies (GWAS), massive amounts of results have been reported on the associations between single-nucleotide polymorphisms (SNPs) and diseases. By now, 2,051 studies and 14, 836 causal variants (p value $\leq$ 5.0×10^{-8}) have been added to catalogue of published Genome-Wide Association Studies [1]. However, it has been found that the effective sizes of the loci identified via GWAS are relatively small and a large proportion of heritability is still missing. This missing heritability problem has been studied by either considering gene-gene and gene-environment interactions or investigating rare variants based on new generation sequencing technology.

Traditional statistical methods are not well suited for detecting such interactions since the number of SNPs and their interactions increase exponentially. To address these issues, many bioinformatics methods for identifying gene-gene interactions have been proposed and one such method is multifactor dimensionality reduction (MDR) [2]. The MDR method is a computationally efficient method for detecting higher-order interactions between genes (and/or gene-environmental factors) and a binary phenotype. The key idea of MDR is to reduce multidimensional genotypes into one-dimensional binary attributes by using a well-defined classifier. Many modifications and extensions of MDR have been developed, which include log-linear models [3], generalized linear models [4], and model-based methods [5]. Among those, the generalized multifactor dimensionality reduction (GMDR) method extends MDR to both dichotomous and continuous phenotypes and allows for the adjustment of covariates such as age, sex, and other clinical variables.

In this study, we focus on gene-gene and/or gene-environment interactions associated with the survival phenotype. In a prospective cohort study, survival time has been one of the important phenotypes in studies of associations with gene expression levels measured by high-throughput microarray technology. Similarly, it has been important to

identify the effect of SNPs on the survival phenotype in GWAS. A series of extensions of MDR to the survival phenotype has recently been proposed, which includes Surv-MDR [6], Cox-MDR [7], and AFT-MDR [8]. Those methods propose new statistics for classifying multilevel genotypes into a binary attribute under the MDR framework. However, as shown in the earlier simulation results [8], Cox-MDR has reasonable power in most cases and is robust to the censoring fraction, while AFT-MDR has similar power as Cox-MDR under no censoring but is very sensitive to the fraction of censoring. It is shown that the power of AFT-MDR substantially reduces, when the fraction of censoring increases more than 30%. That is why we propose two extensions of AFT-MDR, called dAFT-MDR and rAFT-MDR, to improve the power of AFT-MDR under heavier censoring.

Recently, a simple approach to MDR analysis of gene-gene interactions for quantitative traits, called QMDR, has been proposed [9]. The QMDR method replaces the balanced accuracy with a t-test statistic as a score to determine the best interaction model, which yields much less computing load. We extend the idea of quantitative MDR (QMDR) algorithm to Cox-MDR and AFT-MDR methods and propose two extensions of QMDR, called qCox-MDR and qAFT-MDR.

We compare the power of the proposed methods for various parameters including heritability, minor allele frequency (MAF), and censoring proportion with and without adjustment of covariates. It has been found that the improvements of AFT-MDR are less sensitive to censoring fraction than the original AFT-MDR but tend to have less power as the effect of covariate increases. On the other hand, the improvement of Cox-MDR is relatively robust to censoring fraction and tends to have reasonable power across many combinations of parameters.

2. Materials and Methods

2.1. Surv-MDR, Cox-MDR, and AFT-MDR. Since the MDR method has been originally proposed for a binary phenotype in case-control study, it was extended to quantitative traits and various sampling designs. Among those, the Surv-MDR was first proposed [6] for the survival phenotype by using the log-rank test statistic to classify the multi-genotypes into high and low risk groups. It replaces balanced accuracy by log-rank test statistics to determine the best model. However, Surv-MDR cannot allow for covariate adjustment, although adjustment of individual-specific covariates is very important in association studies to remove the confounding effect of covariates.

To overcome the drawback of Surv-MDR, the Cox-MDR method was proposed [7], in which the martingale residual of a Cox model is used as a new score for classifying high and low risk groups. In other words, if the sum of martingale residuals is positive for a specific genotype combination, then the corresponding genotype combination is classified as high risk group or low risk group, otherwise. Once all of genotype combinations are classified as either high or low risk group, the same procedure of original MDR algorithm is implemented to find the best interaction model. Since the martingale residual is obtained from a Cox model with adjusting covariates, the confounding effect of the covariates can be adjusted. It was shown from the simulation result in [7] that Cox-MDR has greater power than Surv-MDR and becomes much better when the effect of covariate increases. Furthermore, Cox-MDR keeps reasonable power even when the fraction of censoring increases, which implies that Cox-MDR is robust to heavier censoring.

Similarly, the AFT-MDR method has also been proposed by using the standardized residual as a new classifier under the accelerated failure time model [8] and the power of AFT-MDR is compared with that of Cox-MDR. As shown in the simulation results [8], the power of Cox-MDR seems to be reasonable in most cases and be robust to the fraction of censoring while the power of AFT-MDR decreases sensitively as the fraction of censoring increases, whereas it has similar power as Cox-MDR under no censoring. From the simulation results, it is shown that the power of AFT-MDR substantially reduces when the fraction of censoring increases more than 30%. Since censoring is very common to occur in survival data, we need to make AFT-MDR more robust to heavier censoring.

2.2. Improvements of AFT-MDR: dAFT-MDR and rAFT-MDR. As mentioned in the previous section, the improvement of AFT-MDR is needed to make it more robust to the fraction of censoring. Based on the simulated data in [8], the distribution of the standardized residual tends to have a long tail as the censoring fraction increases. Then the outliers may have a strong impact on the sum of the standardized residuals in AFT-MDR as those do on the mean value. We consider two different improvements to reduce the effect of the extreme values on the sum of standardized residuals in AFT-MDR.

We first transform the continuous standardized residual into a binary variable instead of taking their sum as done in AFT-MDR. In other words, the individual having the positive standardized residual is regarded as a control, whereas the individual having the negative standardized residual is regarded as a case. As a result, all data is discretized into 0 or 1 and then the original MDR algorithm is implemented, which is called dAFT-MDR (discretized AFT-MDR). Though dAFT-MDR is based on a binary value as the original MDR, it can adjust the covariate effect using the standardized residual of the AFT model, whereas the original MDR cannot adjust the covariate effect.

Secondly, we specify the lower and upper bounds of the standardized residuals and replace the extreme values of the standardized residuals beyond these bounds by either lower or upper bounds. Then we apply the algorithm of AFT-MDR, which is called rAFT-MDR (restricted AFT-MDR). By replacing the extreme values by the prespecified thresholds, the effect of the outliers on the standardized residual may be weakened when the distribution of the standardized residual is extremely skewed under the heavier censoring. However, the determination of threshold of the lower and upper bounds seems to be arbitrary and it should be considered with the behavior of the standardized residuals.

Table 1: The false detection rate of AFT-MDR, dAFT-MDR, rAFT-MDR, Cox-MDR, qCox-MDR, and qAFT-MDR for the log-normal distribution with C_p and MAF when $\gamma = 0$.

MAF	C_p	AFT-MDR	dAFT-MDR	rAFT-MDR	Cox-MDR	qCox-MDR	qAFT-MDR
0.2	0	0.008	0.004	0.006	0.006	0.008	0.003
0.2	0.3	0.002	0.005	0.008	0.006	0.007	0.005
0.2	0.5	0.007	0.005	0.004	0.005	0.004	0.003
0.4	0	0.003	0.006	0.006	0.004	0.008	0.006
0.4	0.3	0.004	0.004	0.005	0.003	0.006	0.003
0.4	0.5	0.007	0.006	0.005	0.006	0.005	0.008

MAF: minor allele frequency; C_p: censoring proportion.

2.3. Improvements of Cox-MDR and AFT-MDR: qCox-MDR and qAFT-MDR. Recently, a simple MDR approach called QMDR for the quantitative trait has been proposed [9], in which the t-test statistic is used to determine the best interaction model in the frame of MDR. The key idea of QMDR can be easily adapted to modify Cox-MDR and AFT-MDR since both the martingale and standardized residuals are quantitative variables.

For Cox-MDR, we obtain the mean value of the martingale residual for each genotype combination and then compare it with the overall mean of the martingale residual. If the mean value of the martingale residual from the specific genotype combination is greater than the overall mean, the corresponding genotype is considered high risk group. Otherwise, it is considered low risk group, since the larger value of martingale residual has higher risk than expected. Once all of the genotypes are classified as high risk and low risk groups, a new binary attribute is created by pooling the high risk genotype combinations into one group and the low risk into another group. Then we use a t-test statistic to test the significant difference between high and low risk groups and choose the best model. The cross validation procedure for QMDR is the same as that used in original MDR. The difference is that the training score and testing score from the t-test statistics are used instead of training and testing balanced accuracies. As done in MDR, the training scores to determine the best k-order interaction model are computed and the maximum testing score is used to identify the best overall model. Similarly, the AFT-MDR method is also improved by using t-test statistic calculated from the standardized residuals of high and low risk groups. These improvements are called qCox-MDR and qAFT-MDR, respectively.

3. Simulation Results

We propose various improvements of AFT-MDR and Cox-MDR to increase the power for detecting gene-gene interactions with the survival phenotype. We implement the comprehensive simulation studies to compare the power of these improvements with those of original AFT-MDR and Cox-MDR.

For the simulation studies, the two disease-causal SNPs are considered among 20 unlinked diallelic loci with the assumption of Hardy-Weinberg equilibrium and linkage equilibrium. For the covariate adjustment, we consider only one covariate which is associated with the survival time but has no interactions with any SNPs. The simulation datasets are generated from different penetrance functions which define a probabilistic relationship between a status of high or low risk groups and SNPs. We consider eight different combinations of two minor allele frequencies of 0.2 and 0.4 and the four different heritabilities of 0.1, 0.2, 0.3, and 0.4. For each of the eight heritability-MAF combinations, a total of 5 models are generated, which yield 40 epistatic models with various penetrance functions, as described in [10].

Suppose that SNP1 and SNP2 are the two disease-causal SNPs and let f_{ij} be an element from the ith row and jth column of a penetrance function. Then we have the following penetrance function:

$$f_{ij} = P\left(\text{high risk} \mid \text{SNP1} = i, \text{SNP2} = j\right). \quad (1)$$

We generate 200 high risk patients and 200 low risk patients for each of the 40 models which depend on the penetrance function, MAF, and heritability. A more detailed description about the heritability assumption is given in [11]. For each dataset, we implement 5-fold cross validation and repeat it 10 times to reduce the fluctuation due to chance of divisions of the data. As a result, we have 100 datasets for each model.

To generate the survival time, we consider three different models: log-normal, Weibull, and Cox model. For each model, the effect size of the genetic factor is fixed as 1.0 and the effect sizes of adjusted covariate are given as $\gamma = 0.0$, 1.0. For the censoring fraction, we consider three different censoring proportions, $C_p = 0.0, 0.3, 0.5$, because the power of AFT-MDR shows substantially decreasing trend when the censoring is heavier than 0.3 in the previous simulation results [8].

First, we check whether the false detection rate is close to the expected value when there is no gene-gene interaction effect because the best model is selected using the maximum balanced accuracy in the algorithm of MDR. To do this, we generate 100 datasets from each of the 40 models, which is a total of 4000 null datasets. Here the false detection rate is estimated as the percentage of times that the method randomly chooses the two disease-causal SNPs as the best model out of each set of 100 datasets for each model. Table 1 shows the false detection rate for AFT-MDR, dAFT-MDR, rAFT-MDR, Cox-MDR, qCox-MDR, and qAFT-MDR for the log-normal distribution when the effect size of the

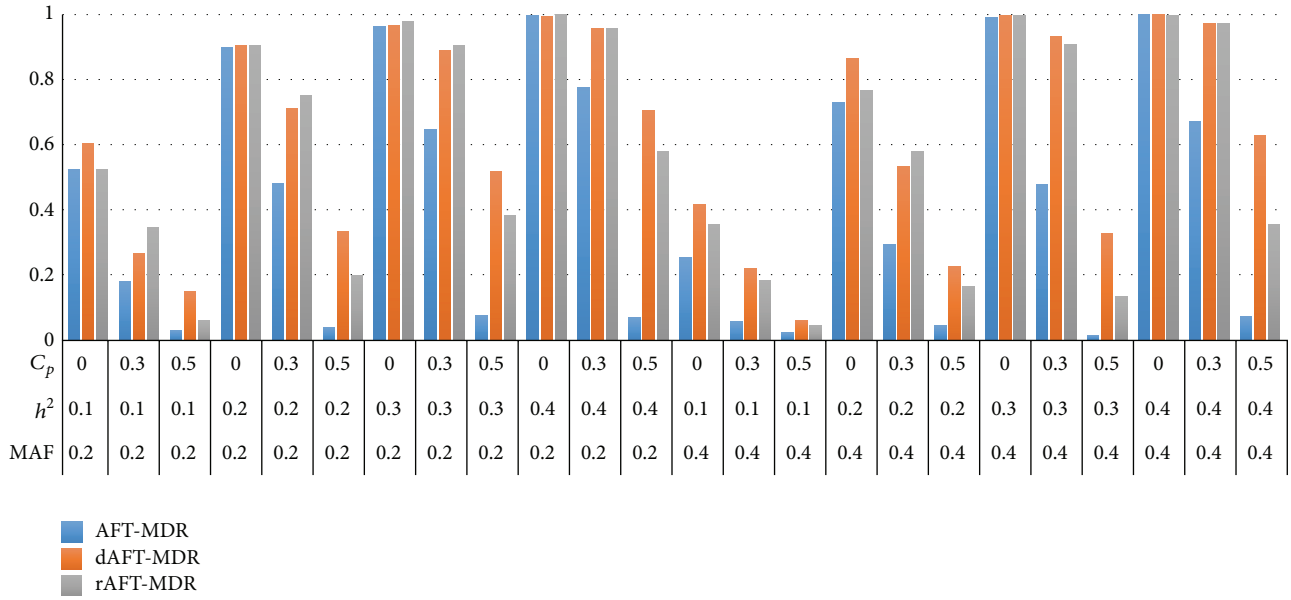


Figure 1: Comparison of the power of AFT-MDR, dAFT-MDR, and rAFT-MDR for the log-normal distribution when $\gamma = 0.0$. *MAF: minor allele frequency; h^2: heritability; C_p: censoring proportion.

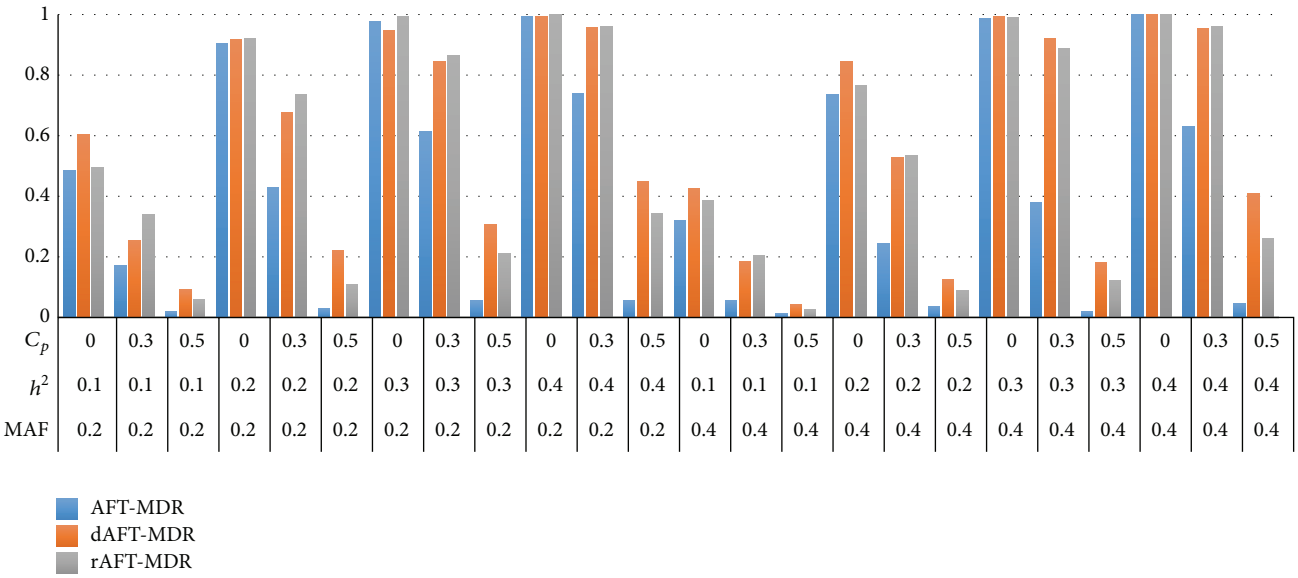


Figure 2: Comparison of the power of AFT-MDR, dAFT-MDR, and rAFT-MDR for the log-normal distribution when $\gamma = 1.0$. *MAF: minor allele frequency; h^2: heritability; C_p: censoring proportion.

adjusting covariate is given as $\gamma = 0.0$. Since only two disease-causal SNPs are considered among 20 SNPs, the expected false detection rate is given as 0.005. As shown in Table 1, the false detection rate varies from 0.002 to 0.008 across the combination of censoring proportion and MAF. For other simulation settings, the false detection rate behaves similarly as shown in Table 1 though not displayed here. It can be concluded that the false detection rate is close to the expected value.

For the power, we consider 100 simulated datasets for each of the 40 models, including two disease-causal SNPs, and we selected the best model over all possible two-way interaction models without and with adjustment of covariates, respectively. The power of dAFT-MDR is estimated as the percentage of times dAFT-MDR correctly chooses the two disease-causal SNPs as the best model out of each set of 100 datasets for each model. The power of the other improvements is defined as the same way of that of dAFT-MDR.

Figures 1 and 2 present the power of AFT-MDR, dAFT-MDR, and rAFT-MDR under the log-normal distribution when $\gamma = 0$ and $\gamma = 1$, respectively. As shown in

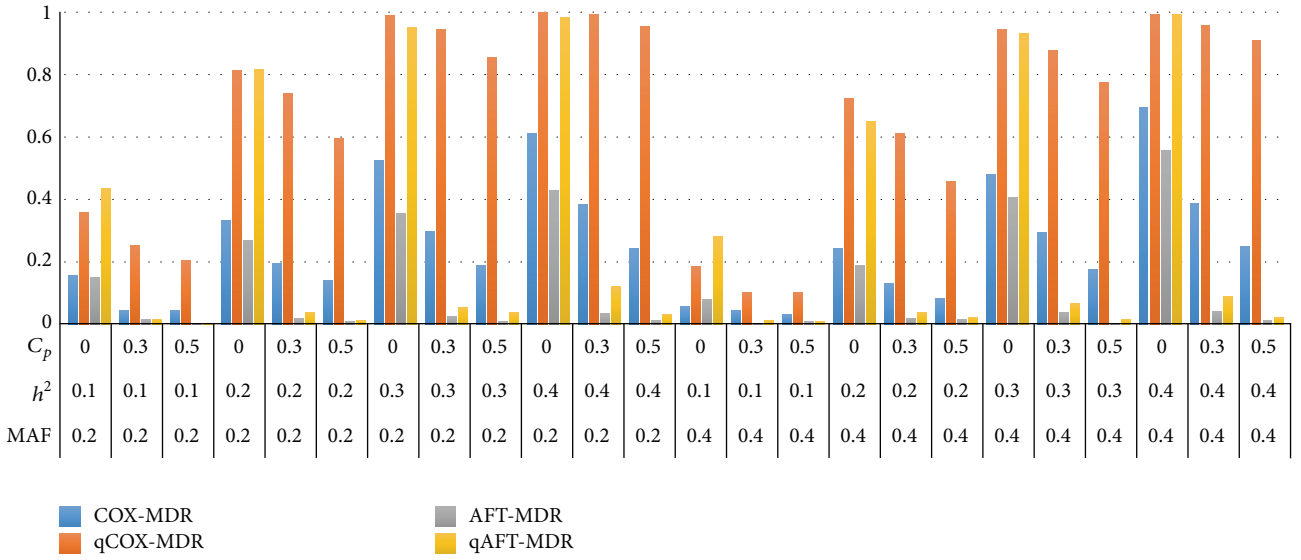


FIGURE 3: Comparison of the power of Cox-MDR, qCox-MDR, AFT-MDR, and qAFT-MDR for a Cox model when $\gamma = 0.0$. *MAF: minor allele frequency; h^2: heritability; C_p: censoring proportion.

Figures 1 and 2, the power of AFT-MDR, dAFT-MDR, and rAFT-MDR has similar trend, which implies that the power of three methods increases as the heritability increases but is lower when the MAF increases from 0.2 to 0.4. As expected, the power of these three methods decreases as the censoring proportion increases from 0.0 to 0.5. In particular, the power of AFT-MDR decreases dramatically when the censoring proportion is lower than 0.3, whereas the power of dAFT-MDR and rAFT-MDR decreases gradually up to the censoring proportion of 0.3 but it decreases faster when the censoring proportion is 0.5. For example, when the MAF is 0.2, heritability is 0.2 and the censoring proportion increases from 0.0 to 0.3 and the power of AFT-MDR decreases from 0.9994 to 0.476 but the power of dAFT-MDR decreases from 0.9904 to 0.8068 and the power of rAFT-MDR decreases from 0.9992 to 0.8072, respectively. Furthermore, when the censoring proportion increases from 0.3 to 0.5, the power of AFT-MDR decreases to 0.0292, whereas the power of dAFT-MDR and rAFT-MDR decreases to 0.3322 and 0.1838, respectively. The degree of decreasing in power is substantially different by improvement in the sense that AFT-MDR hardly detects the significant gene-gene interactions associated with the survival time when the censoring is heavier than 0.5, whereas the improvements of AFT-MDR barely detect the gene-gene interactions. As the heritability increases, the power of AFT-MDR does not increase at all but the power of dAFT-MDR and rAFT-MDR increases up to 0.7026 and 0.5925, respectively. Comparing the power of dAFT-MDR with that of rAFT-MDR, these two improvements seem to behave similarly under the moderate censoring proportion but dAFT-MDR performs better than rAFT-MDR under the heavier censoring as mentioned. This implies that discretizing the standardized residual is more effective than restricting the extreme values as the censoring proportion is heavier than 0.5.

On the other hand, the power of AFT-MDR, dAFT-MDR, and rAFT-MDR behaves similarly when the effect of the covariate increases from $\gamma = 0.0$ to $\gamma = 1.0$ as shown in Figures 1 and 2. This is because the effect of covariate is adjusted by calculating the standardized residual from the AFT model with the adjusted covariates. In addition, the simulation results for the Weibull distribution show the same trend as those for the log-normal distribution though not shown here.

Figures 3 and 4 show the power of Cox-MDR, qCox-MDR, AFT-MDR, and qAFT-MDR for a Cox model and the log-normal distribution, respectively, when the effect size of the adjusted covariate is $\gamma = 0.0$. The power of these four methods performs similarly when the covariate effect is $\gamma = 1.0$. In addition, the power of these four methods for the log-normal distribution is almost the same as that for Weibull distribution though not shown here.

Comparing the simulation results shown in Figures 3 and 4, the power of Cox-MDR, qCox-MDR, AFT-MDR, and qAFT-MDR for a Cox model is rather lower than that for the log-normal model though these two power trends are consistent under the various combinations of the MAF, heritability, and the censoring proportion. The power of these four methods commonly increases as the heritability increases but decreases as the censoring proportion increases and the MAF increases from 0.2 to 0.4. However, the power of Cox-MDR and AFT-MDR is always lower than that of qCox-MDR and qAFT-MDR and decreases substantially as the censoring is heavier than 0.3. For a Cox model, when the MAF is 0.2, the heritability is 0.2 and the censoring proportion increases from 0.0 to 0.3, the power of Cox-MDR decreases from 0.334 to 0.196, and the power of AFT-MDR decreases from 0.270 to 0.018, respectively, whereas the power of qCox-MDR decreases from 0.812 to 0.738 and the power of qAFT-MDR decreases from 0.818 to 0.038, respectively.

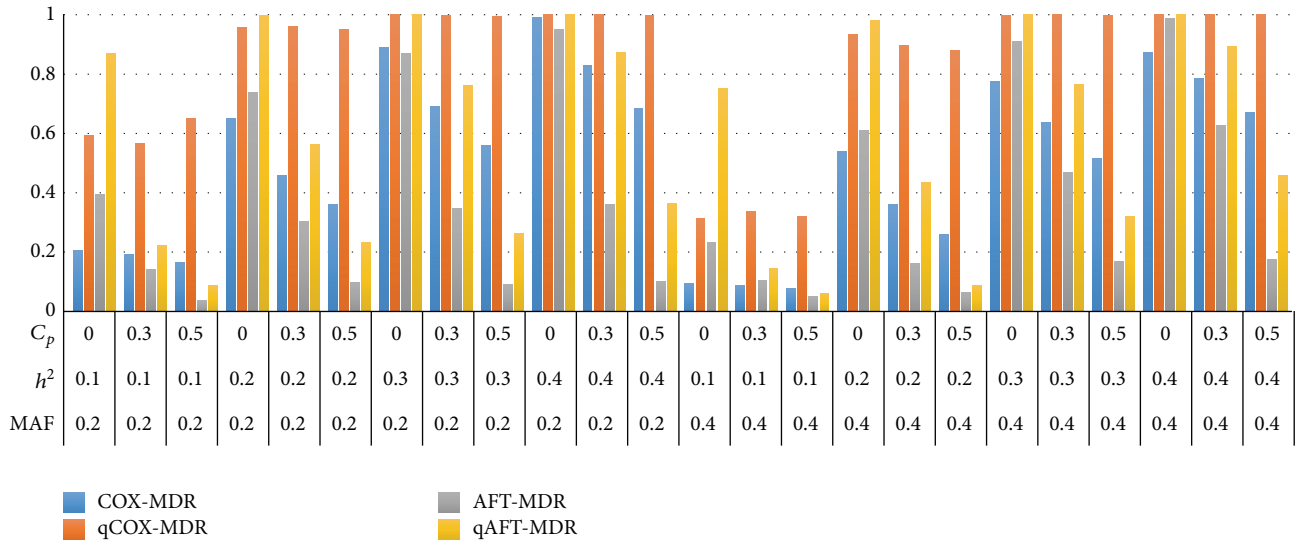


Figure 4: Comparison of the power of Cox-MDR, qCox-MDR, AFT-MDR, and qAFT-MDR for a log-normal distribution when $\gamma = 0.0$. *MAF: minor allele frequency; h^2: heritability; C_p: censoring proportion.

Furthermore, when the censoring is heavier than 0.5, the power of Cox-MDR and AFT-MDR decreases to 0.142 and 0.010, respectively, whereas the power of qCox-MDR and qAFT-MDR decreases to 0.594 and 0.014, respectively. As shown in Figure 3, only the power of qCox-MDR is robust to heavy censoring mechanism, whereas the power of Cox-MDR, AFT-MDR, and qAFT-MDR is very low when the censoring proportion is heavier than 0.3.

On the other hand, for a log-normal model, the power of Cox-MDR decreases from 0.650 to 0.458 as the censoring fraction increases to 0.3 when the MAF is 0.2 and the heritability is 0.2, whereas the power of qCox-MDR changes from 0.958 to 0.960. In addition, the power of Cox-MDR decreases to 0.360 as the censoring fraction increases to 0.5, but the power of qCox-MDR is 0.95, which implies that qCox-MDR is very robust to the censoring fraction. Under the same setting, however, the power of AFT-MDR decreases from 0.738 to 0.302 and the power of qAFT-MDR decreases from 0.998 to 0.564, respectively, as the censoring fraction increases to 0.3. As the censoring fraction increases to 0.5, the power of AFT-MDR and qAFT-MDR decreases to 0.098 and 0.232, respectively. This result is consistent for both the Cox model and the log-normal model, which implies that only the power of qCox-MDR is robust to heavy censoring, though the power of qAFT-MDR is rather higher for the log-normal model than that for Cox model. These trends are similar for Weibull distribution.

In summary, the simulation results show that AFT-MDR, dAFT-MDR, rAFT-MDR, and qAFT-MDR are more sensitive to heavy censoring (more than 0.5) than Cox-MDR and qCox-MDR across various situations. However, for the moderate censoring (less than 0.3), dAFT-MDR, rAFT-MDR, and qAFT-MDR perform much better than the original AFT-MDR.

4. Conclusions

Since many findings from GWAS have been published for the last decades, there is still a missing heritability problem. In order to search the missing heritability, we focus on gene-gene interactions because most of common diseases may be due to the complexity of gene-gene and/or gene-environment interactions rather than a single gene effect. Many plausible approaches have been developed by extending existing methods into a more general framework.

In this paper, we propose various improvements to AFT-MDR and Cox-MDR, which include dAFT-MDR, rAFT-MDR, qAFT-MDR, and qCox-MDR. The motivation to propose dAFT-MDR and rAFT-MDR is to improve the power of AFT-MDR because the performance of AFT-MDR is poor when censoring becomes heavier than 0.3. To reduce the effect of heavy censored observation, we discretize the standardized residual into a binary value, which yields dAFT-MDR. Alternatively, we truncate the extreme values and replace them by specified lower and upper bounds, which leads to rAFT-MDR. As shown in the simulation results, both AFT-MDR and rAFT-MDR have larger powers than the original AFT-MDR for the moderate censoring but still have low powers for the heavy censoring.

In addition, we considered the improvement of QMDR, which has been recently proposed in [9]. By regarding the martingale residual and the standardized residual as the quantitative traits, we adapted the main idea of QMDR and applied it to Cox-MDR and AFT-MDR, which yield qCox-MDR and qAFT-MDR, respectively. As shown in the simulation results, qCox-MDR and qAFT-MDR provided improved performances compared to those of the original Cox-MDR and AFT-MDR, respectively. In particular, qCox-MDR showed the consistent power regardless of the censoring fraction. However, qAFT-MDR yielded the weak power

when the censoring fraction is heavier than 0.3. The censoring fraction seems to have a larger effect on the standardized residual than on the martingale residual. It would be desirable to consider how to make the standardized residual more robust to censoring mechanism.

In conclusion, the improvement of Cox-MDR, say qCox-MDR, has reasonable power and is robust to the heavy censoring, whereas the several improvements of AFT-MDR, say dAFT-MDR, rAFT-MDR, and qAFT-MDR, perform better than AFT-MDR but are not robust to heavy censoring. More studies on the behavior of the standardized residuals are needed to improve the power of AFT-MDR under the heavier censoring.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Education, Science and Technology of Korea (MEST) (NRF: 2013R1A1A3010025 and 2013M3A9C4078158).

References

[1] L. Hindorff, J. MacArthur, J. Morales et al., "A catalog of published Genome-Wide Association Studies," 2014, http://www.genome.gov/GWAStudies/.

[2] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[3] S. Y. Lee, Y. Chung, R. C. Elston, Y. Kim, and T. Park, "Log-linear model-based multifactor dimensionality reduction method to detect gene-gene interactions," *Bioinformatics*, vol. 23, no. 19, pp. 2589–2595, 2007.

[4] X.-Y. Lou, G.-B. Chen, L. Yan et al., "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence," *American Journal of Human Genetics*, vol. 80, no. 6, pp. 1125–1137, 2007.

[5] M. Calle, V. Urrea, N. Malats, and K. van Steen, "MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data," Tech. Rep. 24, Department of Systems Biology, Universitat De Vic, 2008.

[6] J. Gui, J. H. Moore, K. T. Kelsey, C. J. Marsit, M. R. Karagas, and A. S. Andrew, "A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis," *Human Genetics*, vol. 129, no. 1, pp. 101–110, 2011.

[7] S. Y. Lee, M.-S. Kwon, J. M. Oh, and T. Park, "Gene-gene interaction analysis for the survival phenotype based on the cox model," *Bioinformatics*, vol. 28, no. 18, pp. i582–i588, 2012.

[8] J. S. Oh and S. Y. Lee, "An extension of multifactor dimensionality reduction method for detecting gene-gene interactions with the survival time," *Journal of the Korean Data & Information Science Society*, vol. 25, no. 5, pp. 1057–1067, 2014.

[9] J. Gui, J. H. Moore, S. M. Williams et al., "A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits," *PLoS ONE*, vol. 8, no. 6, Article ID e66545, 2013.

[10] D. R. Velez, B. C. White, A. A. Motsinger et al., "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiology*, vol. 31, no. 4, pp. 306–315, 2007.

[11] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *BioData Mining*, vol. 5, article 16, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 671349, 9 pages
http://dx.doi.org/10.1155/2015/671349

Research Article

On the Estimation of Heritability with Family-Based and Population-Based Samples

Youngdoe Kim,[1,2] Young Lee,[1] Sungyoung Lee,[3] Nam Hee Kim,[1] Jeongmin Lim,[4] Young Jin Kim,[1] Ji Hee Oh,[1] Haesook Min,[1] Meehee Lee,[1] Hyeon-Jeong Seo,[1] So-Hyun Lee,[1] Joohon Sung,[5] Nam H. Cho,[6] Bong-Jo Kim,[1] Bok-Ghee Han,[1] Robert C. Elston,[7] Sungho Won,[8] and Juyoung Lee[1]

[1] *The Center for Genome Science, Korea National Institute of Health, KCDC, Osong 361-951, Republic of Korea*
[2] *Department of Applied Statistics, Chung-Ang University, Seoul 156-756, Republic of Korea*
[3] *Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Republic of Korea*
[4] *Chunlab Inc., Seoul National University, Seoul 151-742, Republic of Korea*
[5] *Department of Epidemiology, Seoul National University, Seoul 151-742, Republic of Korea*
[6] *Department of Preventive Medicine, Ajou University School of Medicine, Suwon 443-380, Republic of Korea*
[7] *Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106-7281, USA*
[8] *Department of Epidemiology and Biostatistics, School of Public Health & Institute of Health and Environment, Seoul National University, Seoul 151-742, Republic of Korea*

Correspondence should be addressed to Sungho Won; won1@snu.ac.kr and Juyoung Lee; jylee@cdc.go.kr

Received 22 November 2014; Revised 16 April 2015; Accepted 21 April 2015

Academic Editor: Kristel van Steen

For a family-based sample, the phenotypic variance-covariance matrix can be parameterized to include the variance of a polygenic effect that has then been estimated using a variance component analysis. However, with the advent of large-scale genomic data, the genetic relationship matrix (GRM) can be estimated and can be utilized to parameterize the variance of a polygenic effect for population-based samples. Therefore narrow sense heritability, which is both population and trait specific, can be estimated with both population- and family-based samples. In this study we estimate heritability from both family-based and population-based samples, collected in Korea, and the heritability estimates from the pooled samples were, for height, 0.60; body mass index (BMI), 0.32; log-transformed triglycerides (log TG), 0.24; total cholesterol (TCHL), 0.30; high-density lipoprotein (HDL), 0.38; low-density lipoprotein (LDL), 0.29; systolic blood pressure (SBP), 0.23; and diastolic blood pressure (DBP), 0.24. Furthermore, we found differences in how heritability is estimated—in particular the amount of variance attributable to common environment in twins can be substantial—which indicates heritability estimates should be interpreted with caution.

1. Introduction

Under polygenic inheritance, the effects of segregation at single loci are assumed to be too small to estimate individually and the total genetic variance has been considered to identify the overall genetic effect underlying a trait. Genetic variance consists of additive, dominant, and epistatic components. However, the amount of dominant variance is usually assumed to be relatively small compared to the additive variance and is never identified without a family-based sample that includes bilineal relatives. Similarly, estimation of the epistatic variance (which may include additive components) requires special relationships in family data and is also assumed to be small. Therefore, the estimation of genetic variance has been confined to the additive genetic variance and, to estimate heritability, the proportion of the phenotypic variance attributable to only additive genetic variance has been used even though this can lead to biased estimation in the presence of dominant variance, epistatic variance, and gene × environmental interaction [1].

In general, a parameter allowing for additive polygenic variance can be incorporated into the phenotypic covariances between pairs of individuals, and there are two main ways for incorporating this parameterization. In the absence of population substructure, dominance or any environmental effect shared by family members, the phenotypic covariances can be expressed as a function of the kinship coefficient between family members in family-based samples. Under this parameterization, the additive polygenic variance is obtained from the covariances between family members using variance component models [2–5]. Alternatively, since the advent of large-scale genome data, which reveals similarity in genotypic background, the genetic relationships between individuals have become estimable from genome-wide data and this has also been used to identify population substructure. In the same context, the phenotypic variance explained by additive polygenic variance can also be estimated in population-based samples from the genetic relationships obtained in this way [6, 7]. In particular, the individuals in population-based samples are not closely related and share much less common environmental exposures than do the family members in family-based samples. For this reason Yang et al. [8] suggested excluding closely related individuals from the analysis when estimating heritability from population-based samples, noting that the environmental effect shared by family members seems to be inversely related to their degree of physical proximity, so that close relatives inflate any estimate of heritability.

In this paper, motivated by wishing to calculate the heritability of cardiovascular disease related traits in the Korean population, we examine to what extent estimates of heritability depend on how they are estimated. We calculate the heritability of various traits related to cardiovascular disease in a Korean population using two family-based cohorts, the healthy Twin Study, Korea (HTK) [9] and Ansung Family (ASF) cohorts, and one population-based cohort, that for the Korean Association Resource (KARE) [10] project. Comparing the heritability estimates from family-based and population-based samples, disturbing differences were found. With simulation studies we show that the meaning of heritability estimates can be affected by the absence of highly correlated samples and be substantially inflated by variance attributable to common environment. Thus heritability estimates should be interpreted with caution.

2. Materials

Three cohorts, all part of the Korean Genome Epidemiology Study (KoGES) which is an ongoing prospective epidemiological study, have been utilized to estimate heritability: the KARE project [10] cohort and the HTK [9] and ASF cohorts. These cohorts were genotyped in the Korean Genome Analysis Project (KoGAP) by the Center for Genome Science in the Korea Center for Disease Control and Prevention, which was launched in Korea between 2001 and 2007.

2.1. KARE Project. The KARE project, with 10,038 participants who were living in Ansung (rural) and Ansan (urban), was initiated in 2007 for large-scale genome-wide association studies (GWAS) based on the Korean population. Among the 10,038 participants, 10,004 individuals were genotyped for 500,568 SNPs with the Affymetrix Genome-Wide Human SNP array 5.0. We discarded SNPs with P values for departure from Hardy-Weinberg equilibrium (HWE) less than 10^{-5}, with genotype call rates less than 95%, or minor allele frequencies (MAF) less than 0.01, leaving 350,364 SNPs for subsequent analysis. Individuals with low call rates (<95%, $n = 401$), high heterozygosity (>30%, $n = 11$), gender inconsistencies ($n = 41$), or serious concomitant illness ($n = 101$) were excluded from analysis, along with 601 individuals related or identical whose computed average pairwise identical in state value was higher than that estimated from first-degree relatives of Korean sib-pair samples (>0.8). In total 8,842 individuals were analyzed. In 20 randomly selected duplicate samples, we found that genotype concordance rates exceeded 99.7%, with no single SNP excessively discordant.

2.2. HTK Cohort. The HTK cohort was initiated to identify genetic variation responsible for complex traits as well as the role of the environment in the etiology of complex diseases. Some healthy twins in this cohort were recruited through advertisements in a nationwide newspaper and through posters in about 300 hospitals. Other twin families were selected from the large Korean Genomic Cohort Study of adult individuals and the KoGAP. Then the family members of the selected twins were recruited into this cohort. It should be noted that health status was not considered for sampling. This type of family study can be useful for detecting quantitative trait loci and genetic variations underlying common diseases [11]. Among the 2,473 participants enrolled from April 2005 to December 2008, there are 990 individuals comprising monozygotic (MZ) twins and 234 individuals comprising dizygotic (DZ) twins, and 1861 of these individuals could be genotyped with the Affymetrix Genome-Wide Human SNP array 6.0. We discarded SNPs with P values for departure from HWE less than 10^{-5} or MAF less than 0.01. In addition, SNPs were excluded if Mendelian errors or double recombinants were found in at least 3 families, and in total 520,484 SNPs were used for analysis. We calculated the proportion of genotypes identical in state between individuals in each family and excluded those with any inconsistency between the genetic and reported relationship ($n = 58$). Also, individuals who had coding errors for MZ/DZ status ($n = 2$) were excluded, and as a result genotypes for 1801 family members were available for analysis. Among the genotyped individuals, there are 4 pairs of MZ twins and 393 genotyped individuals whose MZ twin siblings were not genotyped. Also 84 pairs of DZ twins were genotyped, and there are 16 additional genotyped individuals whose DZ twin siblings' genotypes were unknown. There are 162 nuclear families and 3 families consisting of individuals in three generations that include MZ/DZ twins.

2.3. ASF Cohort. In the Ansung area, 5,018 unrelated and related participants were initially recruited for the KARE project; another cohort to study type 2 diabetes was initiated in this area in 2007. In this cohort, some individuals were selected from the KARE project, and their family members and other individuals from the Ansung area who were not in

the KARE project were included, if they were diagnosed as having type 2 diabetes and agreed to participate in this study. This sampling scheme could lead to the presence of ascertainment bias, but the small correlations between type 2 diabetes status and the traits of interest (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/671349) reveal that any ascertainment bias would not be substantial. In these samples, 456 individuals who were included in the KARE project were genotyped with the Affymetrix Genome-wide Human SNP array 5.0, and another 781 individuals were genotyped with the Affymetrix Genome-wide Human SNP array 6.0. Individuals were excluded if they reported relationships in the family inconsistent with the genotypic relationships estimated by the proportion of genotypes identical in state ($n = 41$) or had unavailable trait data ($n = 412$). Also, SNPs were excluded if Mendelian inconsistency was found in at least 3 families, the P values for HWE were less than 10^{-5}, or the MAF was less than 0.01. As a result, 784 family members with 417,719 SNPs were used for our analysis.

3. Methods

To estimate heritability we used the freely available software Genome-wide Complex Trait Analysis (GCTA) [8] and the ASSOC program in the Statistical Analysis for Genetic Epidemiology (S.A.G.E.) [12] package. We considered eight traits: height, body mass index (BMI), triglycerides (TG), total cholesterol (TCHL), high-density lipoprotein (HDL), low-density lipoprotein (LDL), systolic blood pressure (SBP), and diastolic blood pressure (DBP). We included age, age^2, and sex as covariates. In particular, the linear mixed model for GCTA is robust to population substructure, and the EIGENSTRAT method [13] which includes PC scores as covariates was not applied. The effect of a living environment variable (urban versus rural) was not significant at the 0.05 significance level for any of the eight traits and so was not included as a covariate in the detailed analyses reported in Tables S2–S9. Quantile-quantile plots in Figures S1-S2 indicate that TG is not normally distributed and log-transformed TG (log TG) was used to obtain approximate normality. For the other phenotypes, the original scales were used because heritability estimates on the original scale and after inverse-normal transformation were almost the same and interpretation is not straightforward for the inverse-normal transformed data. The missing rates for each phenotype were calculated (see Table S10) and were usually very small. ASSOC parameterizes the phenotypic correlations between individuals using the reported familial relationships and can split the nonpolygenic variance into components for measurement error, sibling, and marital effects, and these results were summarized. GCTA estimates heritability by parameterizing phenotypic correlations with the estimated genetic relationship matrix (GRM) from the standardized genotypes. In particular, the results from GCTA were obtained with and without the default GRM-cutoff option. In addition, we separately analyzed monozygotic (MZ) and dizygotic (DZ) twin data, to estimate the relative proportion of the phenotypic variance attributable to common environmental effects.

3.1. Heritability Estimation Using Familial Relationships. Under the multivariate normality model, the covariance between family members can be expressed as a function of their kinship coefficients and this can be utilized to estimate heritability. We estimated the heritability from the family data, separately in the HTK and ASF cohorts, with the ASSOC program in S.A.G.E. (ver. 6.2) [12]. ASSOC is based on a linear mixed model and the parameters are estimated by the maximum likelihood (ML) method. Let y_{ij} denote the response for individual j in family i, where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$; n and n_i indicate the number of families and the number of individuals in family i, respectively. Also, let x_{ij} indicate covariates that affect y_{ij}. Then, denoting $\pi_{ijj'}$ as the kinship coefficient between individual j and individual j' in family i, we let

$$\mathbf{X}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{pmatrix},$$

$$\mathbf{Y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, \tag{1}$$

$$\mathbf{\Phi}_i = \begin{pmatrix} 1 & 2\pi_{i12} & \cdots \\ 2\pi_{i21} & 1 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}.$$

We denote the additive polygenic, dominant polygenic, and random error variances, respectively, by σ_a^2, σ_d^2, and σ^2. If we also denote the $w \times w$ identity matrix by $\mathbf{I}_w$, the linear model used in ASSOC for random mating and only additive effects is

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad \text{where } \boldsymbol{\varepsilon}_i \sim \text{MVN}\left(0, \sigma^2\mathbf{I}_{n_i} + \sigma_a^2\mathbf{\Phi}_i\right). \tag{2}$$

$\mathbf{\Phi}_i$ will be called the familial relationship matrix (FRM) in the remainder of this paper. Furthermore, ASSOC can estimate the variances separately attributable to polygenic, common sibship, and marital effects as described by Elston et al. [14].

S.A.G.E. ASSOC was used to estimate heritability in the family-based HTK and ASF cohorts and, for a fair comparison with GCTA, only genotyped individuals were analyzed this way. In the HTK cohort, 1801 genotyped individuals were considered, and there are 4 pairs of MZ twins among those genotyped. S.A.G.E. cannot easily handle MZ twins, and a single individual for each MZ twin was randomly selected for analysis with both ASSOC and GCTA. There is other software available that can handle MZ twins [15–17] in pedigrees, but this was not considered because the number of genotyped MZ twins is very small and so the variance attributable to common environment could not be well estimated in these

cohorts using the GRM. We used the program PEDINFO in S.A.G.E. to provide descriptive statistics of the pedigree data.

3.2. Heritability Estimation Using Estimated Genetic Relationships. When large-scale genotypes are available, the GRM can be estimated with the software GCTA [8] and, instead of the FRM, the estimated GRM can be incorporated into the same linear mixed model (2) as available in ASSOC, to estimate σ_a^2. The minor allele frequencies for GRM were estimated by using all individuals even when some individuals were correlated. Because the genetic relationship is estimated with genotypes, GCTA can be applied to both family-based and population-based samples. In addition, the GCTA program can estimate the variance components by both the restricted maximum likelihood (REML) and ML methods. The REML method provides more unbiased estimates of the variance components than the ML method. Therefore we estimated heritability by the REML method when applying GCTA to the KARE project, the HTK cohort, and the ASF cohort, though for these large samples the difference in the estimates is expected to be trivially small. Yang et al. [8] suggested excluding closely related individuals from the analysis when estimating genetic variation captured by all the SNPs, using a GRM-cutoff option. However, for the analysis of family-based samples, family members are highly correlated and most individuals become excluded from the analysis if the GRM-cutoff option for individual selection is activated. We report the results of both with and without the GRM-cutoff option, and we used 0.025 as the GRM-cutoff.

3.3. Estimating Familial Correlations with S.A.G.E. FCOR in S.A.G.E. [12] can estimate familial correlations for all pair types existing in a set of pedigrees. FCOR cannot handle the effect of covariates, and thus for height, BMI, log TG, TCHL, HDL, LDL, SBP, and DBP, we calculated the residuals from the linear model with age, age^2, and sex as covariates. Residuals from this linear model were used to estimate the empirical correlations between family members and their 95% confidence intervals with FCOR in S.A.G.E.

3.4. Estimating Variance Attributable to the Common Environment with Twins. If we assume an additive model with no interaction, the phenotypic variance consists of the genetic variance and a common environmental variance component. However, the variance for environmental effects shared by family members is in general unidentifiable. If we further assume that the amount of covariance between MZ twins attributable to a common environmental effect is similar to that between DZ twins [18] and that any dominant or epistatic polygenic effects are relatively small compared to the additive genetic and common environmental effects, the covariance attributable to the common environmental effect can be estimated.

We separated out all the MZ and DZ twins, whether genotyped or not, from the HTK cohort, so that the members in each family are always either MZ or DZ twins in this analysis. In total, 958 individuals (479 pairs) comprising MZ twins and 224 individuals (112 pairs) comprising DZ twins were analyzed. If we denote the common environmental variance by σ_c^2, the polygenic model provides the following variance-covariance structure between twins:

$$\operatorname{cov}\left(y_{i1}, y_{i2}\right) = \begin{cases} \sigma_c^2 + \sigma_a^2 + \sigma_d^2 & \text{for MZ twins} \\ \sigma_c^2 + 0.5\sigma_a^2 + 0.25\sigma_d^2 & \text{for DZ twins.} \end{cases} \quad (3)$$

To construct this variance-covariance structure for MZ and DZ twins in our linear mixed model, we denote $\sigma_Y^2 = \operatorname{var}(y_{ij})$, $r_{\mathrm{DZ}} = (\sigma_c^2 + 0.5\sigma_a^2 + 0.25\sigma_d^2)/\sigma_Y^2$, and $r_{\mathrm{MZ}} = (\sigma_c^2 + \sigma_a^2 + \sigma_d^2)/\sigma_Y^2$. We define two matrices $\mathbf{A}$ and $\mathbf{B}$ as follows:

$$\mathbf{A} = \left(a_{ij}\right), \quad a_{ij} = \begin{cases} 1 & i, j \text{ are MZ twins} \\ 0 & \text{otherwise,} \end{cases}$$
$$\mathbf{B} = \left(b_{ij}\right), \quad b_{ij} = \begin{cases} 1 & \text{DZ twins} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Then, our linear model becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$
$$\boldsymbol{\varepsilon} \sim \mathrm{MVN}\left(0, \sigma_Y^2\mathbf{V}\right), \text{ where } \mathbf{V} = \mathbf{I}_n + r_{\mathrm{MZ}}\mathbf{A} + r_{\mathrm{DZ}}\mathbf{B}. \quad (5)$$

Here r_{MZ} and r_{DZ} should be between -1 and 1. We used the REML method to estimate variance parameters, and each parameter was estimated by the average information method [19, 20]. R code for the proposed method can be downloaded from http://healthstat.snu.ac.kr/data/heritability_Rcode.zip. It is simple to show that, ignoring any epistatic effects, $2\sigma_{\mathrm{DZ}} - \sigma_{\mathrm{MZ}}$ is $\sigma_c^2 - 0.5\sigma_d^2$ and, if we assume that $\sigma_d^2 = 0$, $2\sigma_{\mathrm{DZ}} - \sigma_{\mathrm{MZ}}$ becomes σ_c^2 and the proportion of variance attributable to common environment, ρ_c, can be calculated by $2r_{\mathrm{DZ}} - r_{\mathrm{MZ}}$. If we let $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}$, the Fisher information matrix for σ^2, r_{MZ}, and r_{DZ} can be obtained by

$$\boldsymbol{\Psi} = \begin{pmatrix} \dfrac{n-p}{2\sigma^4} & \dfrac{1}{2\sigma^2}\operatorname{tr}(\mathbf{PA}) & \dfrac{1}{2\sigma^2}\operatorname{tr}(\mathbf{PB}) \\ \dfrac{1}{2\sigma^2}\operatorname{tr}(\mathbf{PA}) & \dfrac{1}{2}\operatorname{tr}(\mathbf{PAPA}) & \dfrac{1}{2}\operatorname{tr}(\mathbf{PAPB}) \\ \dfrac{1}{2\sigma^2}\operatorname{tr}(\mathbf{PB}) & \dfrac{1}{2}\operatorname{tr}(\mathbf{PAPB}) & \dfrac{1}{2}\operatorname{tr}(\mathbf{PBPB}) \end{pmatrix}, \quad (6)$$

where n is a sample size and p is the number of covariates. Thus the variance of ρ_c can be obtained by $(0, -1, 2)\boldsymbol{\Psi}^{-1}(0, -1, 2)^t$. Provided the environmental correlation is the same for both MZ and DZ twins, this estimate can be utilized as a lower bound for the variance attributable to the environmental effects shared by siblings.

3.5. Simulation Studies. With extensive simulation studies, we investigated the accuracy of heritability estimates for various scenarios. We generated 5000 pairs of individuals with 100,000 SNPs, and heritability was estimated by GCTA without activating the GRM-cutoff. The individuals in different pairs were generated to be independent and the correlations of genotypes, r, between individuals in each pair were generated to be $1/2, 1/4, \ldots, 1/128$, or 0. A pair of individuals

Table 1: Descriptive statistics for the traits in each cohort.

Trait	HTK (n = 1801)			ASF (n = 784)			KARE (n = 8842)		
	1st Q	Median	3rd Q	1st Q	Median	3rd Q	1st Q	Median	3rd Q
Sex (m/f)	711/1090 (0.39/0.61)			372/412 (0.47/0.53)			4183/4659 (0.47/0.653)		
Age	35	43	57	35	46.5	60	44	50	60
Height	155.6	160.9	167.7	155.4	162.1	169	153.3	159.7	166.6
BMI	21.47	23.61	25.9	22.21	24.39	26.64	22.51	24.48	26.5
log TG	4.17	4.55	4.94	4.36	4.71	5.106	4.605	4.913	5.252
HDL	41	48	57	37	43	51	37	44	50
LDL	91	110	132	93.6	115	135.3	114.2	115.7	136.4
SBP	108	118.7	130	110	120	130	104.67	115.33	128
DBP	70	72	80	72	79	84	68.67	74	81.33
TCHL	164	187	211	165	185	210.2	167	189	214

Table 2: Estimates (s.e.) of heritability.

		Cohort			
		Family-based		Population-based	All
		HTK	ASF	KARE	
Traits	Height	0.76 (0.04)	0.66 (0.09)	0.32 (0.04)	0.60 (0.02)
	BMI	0.43 (0.05)	0.41 (0.08)	0.15 (0.04)	0.32 (0.02)
	TG	0.37 (0.05)	0.27 (0.08)	0.21 (0.04)	0.24 (0.02)
	TCHL	0.47 (0.05)	0.50 (0.08)	0.18 (0.04)	0.30 (0.02)
	HDL	0.72 (0.04)	0.50 (0.07)	0.16 (0.04)	0.38 (0.02)
	LDL	0.43 (0.05)	0.47 (0.08)	0.16 (0.04)	0.29 (0.02)
	SBP	0.37 (0.05)	0.23 (0.08)	0.26 (0.04)	0.23 (0.02)
	DBP	0.53 (0.05)	0.21 (0.08)	0.21 (0.04)	0.24 (0.02)

with $r = 1/2$ indicates siblings or a parent-offspring pair. To generate pairs of individuals with correlation of genotypes r, randomly selected alleles from two individuals were generated to be identical by descent with probability $r/2$. The minor allele frequencies were generated from $U(0, 0.4)$ and genotypes were generated with the binomial distribution under Hardy-Weinberg equilibrium (HWE). Monomorphic variants were excluded from the analyses, and all markers were assumed to be in linkage equilibrium. If there are too many redundant SNPs in linkage disequilibrium with the causal variants, the empirical standard deviation of heritability estimates can be inflated and the analysis with GCTA should be modified as indicated by Speed et al. [21].

The traits were generated by summing a polygenic effect and a random effect. The random effect was generated from $N(0, \sigma^2)$. To create a polygenic effect we simulated 100 independent causal SNPs and we assumed that all or 50 randomly selected ones of these causal SNPs were genotyped. The additive disease mode of inheritance was assumed and a single SNP genetic effect is denoted by β_l. Letting p_l be the allele frequency for causal SNP l ($l = 1, 2, \ldots, 100$) and heritability be h^2, the genetic effect, β_l, was calculated as

$$\beta_l = \sqrt{\frac{h^2\sigma^2}{200p_l\left(1-p_l\right)\left(1-h^2\right)}}. \quad (7)$$

Here p_l were generated from $U(0, 0.1)$ or $U(0.1, 0.4)$, respectively, and the genetic effects $2\beta_l^2 p_l(1 - p_l)$, for the 100 causal SNPs, were taken to be equal. σ^2 was assumed to be 1 and h^2 was taken to be 0.1, 0.3, 0.5, 0.7, or 0.9.

4. Results

4.1. Estimates of Heritability in a Korean Population. Table 1 shows the descriptive statistics for eight traits: height, BMI, log TG, TCHL, HDL, LDL, SBP, and DBP. Interquartile ranges for these traits show that the traits in the three cohorts are comparable. We calculated heritabilities in the HTK, ASF, and KARE cohorts separately, and they were also combined to calculate overall heritabilities by pooling the samples and including two dummy (0/1) covariates to adjust for the effects of each sample. Table 2 shows that the heritability estimates from the pooled samples with GCTA were, for height, 0.60; BMI, 0.32; log TG, 0.24; TCHL, 0.30; HDL, 0.38; LDL, 0.29; SBP, 0.23; and DBP, 0.24. In each case these heritability estimates are between the limits of those from the individual KARE, HTK, and ASF cohorts. Tables S2–S9 show that, in the samples where both GCTA and S.A.G.E. can be applied, the heritability estimates from S.A.G.E. and GCTA are usually comparable. GCTA estimates heritability with the REML method based on an estimated GRM, while S.A.G.E. estimates heritability with the ML method based on the FRM. The estimates from the REML and ML methods must be very similar for a large sample size, and thus the convergence of the estimated GRM to FRM [22] explains their similarity.

Table 3: Estimates of variance components in the HTK cohort. MZ and DZ twins were separated out and used to estimate correlations of MZ and DZ twins. ρ_c indicates a lower bound for the proportion of variance explained by the environmental effects shared by family members.

	cor (MZ)[a]	cor (DZ)[b]	$\rho_{c\ (95\%\ \text{confidence interval})}$
Height	0.970	0.832	0.694 (0.690, 0.698)
BMI	0.729	0.232	0.266 (0.061, 0.496)
log TG	0.551	0.336	0.121 (0.070, 0.172)
TCHL	0.624	0.382	0.139 (0.093, 0.185)
HDL	0.677	0.476	0.275 (0.240, 0.310)
LDL	0.656	0.342	0.028 (−0.022, 0.078)
SBP	0.601	0.453	0.306 (0.268, 0.344)
DBP	0.646	0.585	0.524 (0.501, 0.547)

[a]Correlation of MZ twins. [b]Correlation of DZ twins.

4.2. Overestimation of Heritability in Family-Based Samples. From Table 2, we see substantial differences between the heritability estimates from population-based samples and those from family-based samples. Our estimates with population-based samples, KARE, are, for height, 0.32; BMI, 0.15; log TG, 0.21; TCHL, 0.18; HDL, 0.16; LDL, 0.16; SBP, 0.26; and DBP, 0.21 (with the living area variable (urban/rural) included as a covariate; for height, 0.32; BMI, 0.15; log TG, 0.24; TCHL, 0.15; HDL, 0.16; LDL, 0.13; SBP, 0.22; and DBP, 0.16). The largest difference between the family-based and population-based samples was found for HDL, followed by height. However, the phenotypic variances are usually similar and so it is unlikely there exists heterogeneity of heritability between the two types of sample. (It should be noted that the probands in ASF were selected from KARE.) Alternatively, these differences could be explained by the different properties of family-based and population-based samples. The variance attributable to the shared environmental effects by family members was estimated for HTK and ASF with ASSOC. Significant marital effects were found for height and DBP, which, respectively, explain 17% and 16% of the phenotypic variance in the HTK cohort (Tables S2–S9). The marital effect may be related to natural/positive/negative selection and, in particular, assortative mating is known to occur for height [23, 24]. In addition, we found significant common sibling effects (% total phenotypic variance) for BMI, 0.94 (10%); log TG, 0.03 (11%); TCHL, 151.28 (12%); HDL, 9.63 (7%); LDL, 141.50 (16%); and SBP, 23.27 (10%) in the HTK cohort; and HDL, 17.57 (16%) in the ASF cohort (Tables S2–S9). These significantly large percentages indicate a tendency for the environmental elements common to siblings to be similar.

However, even though ASSOC can detect the presence of some environmental effects shared by family members, the heritability estimates it produces for family-based samples are still much larger than those produced by GCTA from population-based samples. Examination of the familial correlations (Table S11) provides evidence that heritability estimated with family-based samples may be inflated if, unlike the analysis we performed with ASSOC, the sibling and marital correlations are ignored. First, the mother-father correlations for height, BMI, and DBP are significantly larger than 0 at the 0.05 significance level, whereas the usual polygenic model assumes that their correlations are 0. At the same time, in large pedigrees this positive mother-father correlation could lead to inflated parent-offspring correlations, but this effect cannot be completely handled in the existing software. Even though ASSOC can allow for a mother-father correlation, the parent-offspring correlation could be larger than expected as a result of the positive mother-father correlation; to allow completely for this, the polygenic variance should be allowed to decrease from one generation to the next. The larger correlations between siblings than those between parents and offspring could thus conceivably be partially attributable to this. Second, correlations between siblings are much larger than those between parents and offspring. In particular for log TG, TCHL, and LDL, this occurs even though the mother-father correlations are around 0. If we assume that dominant polygenic effects are small, the environmental effect shared by siblings seems to be larger than that shared by parents and offspring. The program ASSOC in S.A.G.E. appropriately allows for both a marital correlation and a common sibling component of variance over and above that due to an additive polygenic variance, though that variance is assumed to be constant across generations.

Table 3 shows correlations between DZ and MZ twins that were estimated with the linear mixed model. The correlations between MZ twins are expected to be around twice as large as those between DZ twins in the absence of both environmental effects shared by family members and dominant polygenic effects. However, for all traits other than BMI, twice the correlation between DZ twins is much larger than the correlation between MZ twins. ρ_c shows that the proportion of variance explained by shared environment for height may be 69.4%, and we can conclude that the correlations generated by the environmental effects shared by family members are usually much more substantial than we expect.

4.3. Underestimation of Heritability in Population-Based Samples. Figures 1-2 show heritability with GCTA using the GRM estimated from 100 K simulated SNPs. All causal variants were generated from $U(0, 0.1)$ or $U(0.1, 0.4)$, respectively. Each case was summarized with 200 replicates, and in Figures 1-2, we assumed that the number of causal SNPs was 100 and h^2 was set at 0.5. The results show that heritability estimates are always around the proportion of variances explained by all causal variants, 0.5, when all the causal SNPs are used to

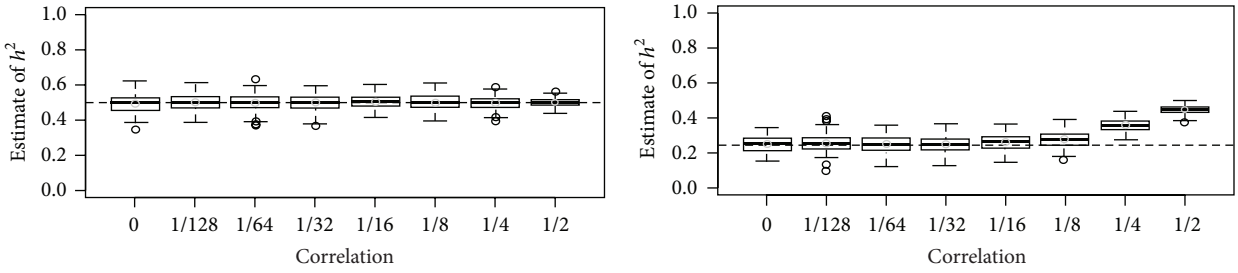


Figure 1: Heritability estimates for various levels of genetic correlation with 10,000 individuals when h^2 was set at 0.5 and all causal variants were generated from $U(0, 0.1)$. We generated 5,000 pairs of individuals with 100,000 SNPs, and each box-plot was generated with results from 200 replicates. The dashed horizontal line indicates the proportion of the total phenotypic variance explained by the SNPs used for calculating the GRM, and the estimates of heritability with GCTA are plotted against the correlation between family members. In (a), 100 causal SNPs were used to estimate the GRM, and in (b), 50 randomly selected causal SNPs were used. The horizontal dotted line indicates the relative proportion of variance explained by the SNPs.

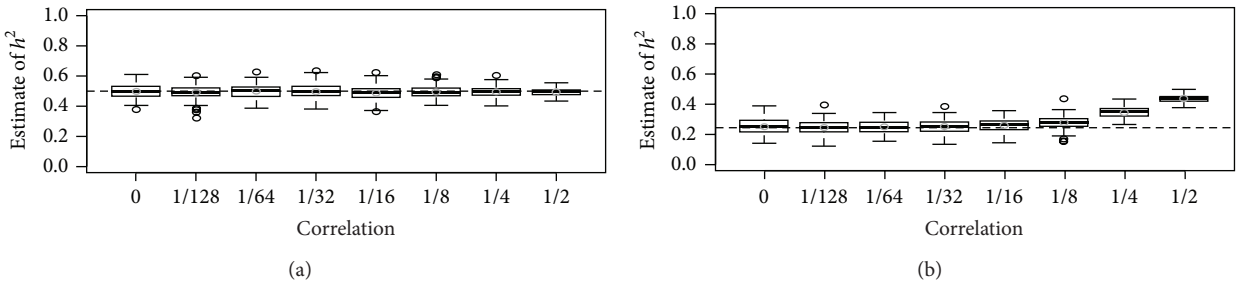


Figure 2: Heritability estimates for various levels of genetic correlation with 10,000 individuals when h^2 was set at 0.5 and all causal variants were generated from $U(0.1, 0.4)$. We generated 5,000 pairs of individuals with 100,000 SNPs, and each box-plot was generated with results from 200 replicates. The dashed horizontal line indicates the proportion of the total phenotypic variance explained by the SNPs used for calculating the GRM, and the estimates of heritability with GCTA are plotted against the correlation between family members. In (a), 100 causal SNPs were used to estimate the GRM, and in (b), 50 randomly selected causal SNPs were used. The horizontal dotted line indicates the relative proportion of variance explained by the SNPs.

estimate the GRM. However, when half the causal SNPs are used to estimate the GRM and r is larger than 0.125, heritability estimates are overestimated. There is a tendency for the overestimation to be proportional to r. Figures 1-2 also show that the interquartile distance for a heritability estimate is inversely related to r. In Supplementary Figures 3–6 h^2 was assumed to be 0.1, 0.3, 0.7, or 0.9, respectively, and we found that our results are the same as in Figures 1-2.

5. Discussion

As a simple dimensionless overall measure of the importance of genetic factors, heritability has been used to determine the potential for predicting the genetic risk of disease. Estimating heritability requires information about genetic or familial relationships to parameterize the variance component explained by genetic factors, and formerly this was feasible only with family-based samples. With the advance of genotyping technology, large-scale genome-wide data has enabled estimation of the GRM from population-based samples, and now both family-based and population-based samples can be utilized to estimate heritability.

Heritability is a population-specific and trait-specific parameter, so it is natural that estimates have been diverse, depending on the samples and traits studied. However, we found substantial differences between the heritability estimates from population-based samples and those from family-based samples for the same trait even though both came from the same country. Although the significant differences between the two heritability estimates might be explained by heterogeneity between the samples, the estimates using population-based samples must be understood as the relative proportion of variance explained by the SNPs used to estimate the GRM [8], and this fact has been utilized to explain the missing heritability. Unbiased heritability estimation requires some individuals with large genotype correlations and the degree of genetic relationship between the individuals studied can be a more influential factor when estimating heritability. Furthermore, we attempted to quantify the variance attributable to common environment with MZ/DZ twins and found that the amount of heritability inflation can be substantial. For instance, the proportion of variance generated by shared environment is 69.4% for height, which indicates that the large value of previously reported heritability

estimates for height may be generated by a large common environment component. If extended families are utilized, the amount of overestimation seems to be less substantial, but further investigation of appropriate statistical methods and study design is necessary on how to prevent the inflation of heritability estimates due to common environment effects.

Of course, in spite of our comprehensive analysis, there are several limitations to our conclusions. First, we estimated the amount of variance attributable to common environment by assuming its equivalence between MZ and DZ twins which, depending on the trait, may not be true; and there may be heteroscedasticity between MZ and DZ twins, or between twins and nontwins. If we have available MZ twins who lived apart, more accurate estimates for the variance attributable to common environment may be obtainable. Second, phenotypic differences between populations can be induced by genetic and/or environmental differences, and under population substructure the phenotypic covariance can be inflated if there are phenotypic differences between populations attributable to environmental differences. Third, it has been shown that epistasis can inflate the additive polygenic variance [1] but it is unclear whether our conclusions are still preserved in such cases. Further studies for better study design and statistical algorithms are necessary to clarify these issues.

Heritability has been a useful measure to motivate genetic studies and many statistical algorithms have been implemented to estimate it. However, complex traits result from a complex interplay of genotype and environment, and any model used to estimate heritability has a limited meaning because of the so-called phantom heritability [1]. Therefore we can conclude that it may not be always good to trust current estimates under the study designs and methodologies employed so far.

6. Conclusion

We estimated the heritability of traits related to cardiovascular disease, from both family-based and population-based samples, collected in Korea, and substantial differences were found between the family-based and population-based samples when using genetic markers to estimate relationship. With extensive simulations, we found that the meaning of heritability estimates can be different depending on the correlations between individuals. Furthermore, we identified the amount of variance attributable to common environment with twins and found that heritability inflation can be substantial, which indicates heritability estimates should be interpreted with caution.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Youngdoe Kim and Young Lee contributed equally to this work.

Acknowledgments

This work was supported by an intramural grant from the Korean National Institute of Health (2011-N73001-00, 2013-NG73002-00) and grants from the Korean Centers for Disease Control and Prevention (4845-301, 4851-302, and 4851-307).

References

[1] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 4, pp. 1193–1198, 2012.

[2] L. Almasy and J. Blangero, "Multipoint quantitative-trait linkage analysis in general pedigrees," *The American Journal of Human Genetics*, vol. 62, no. 5, pp. 1198–1211, 1998.

[3] D. S. Falconer, *Introduction to Quantitative Genetics*, Ronald Press, New York, NY, USA, 1960.

[4] R. A. Fisher, "The correlation between relatives on the supposition of mendelian inheritance," *Transactions of the Royal Society of Edinburgh*, vol. 52, pp. 399–433, 1918.

[5] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era—concepts and misconceptions," *Nature Reviews Genetics*, vol. 9, no. 4, pp. 255–266, 2008.

[6] J. Yang, B. Benyamin, B. P. McEvoy et al., "Common SNPs explain a large proportion of the heritability for human height," *Nature Genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[7] J. Yang, T. A. Manolio, L. R. Pasquale et al., "Genome partitioning of genetic variation for complex traits using common SNPs," *Nature Genetics*, vol. 43, no. 6, pp. 519–525, 2011.

[8] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *American Journal of Human Genetics*, vol. 88, no. 1, pp. 76–82, 2011.

[9] J. Sung, S.-I. Cho, K. Lee et al., "Healthy twin: a twin-family study of Korea—protocols and current status," *Twin Research and Human Genetics*, vol. 9, no. 6, pp. 844–848, 2006.

[10] Y. S. Cho, M. J. Go, Y. J. Kim et al., "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits," *Nature Genetics*, vol. 41, no. 5, pp. 527–534, 2009.

[11] L. Peltonen and V. A. McKusick, "Genomics and medicine: dissecting human disease in the postgenomic era," *Science*, vol. 291, no. 5507, pp. 1224–1229, 2001.

[12] R. C. Elston and C. Gray-McGuire, "A review of the 'Statistical Analysis for Genetic Epidemiology' (S.A.G.E.) software package," *Human genomics*, vol. 1, no. 6, pp. 456–459, 2004.

[13] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[14] R. C. Elston, V. T. George, and F. Severtson, "The Elston-Stewart algorithm for continuous genotypes and environmental factors," *Human Heredity*, vol. 42, no. 1, pp. 16–27, 1992.

[15] M. Neijts, J. van Dongen, C. Kluft, D. I. Boomsma, G. Willemsen, and E. J. de Geus, "Genetic architecture of the pro-inflammatory state in an extended twin-family design," *Twin Research and Human Genetics*, vol. 16, no. 5, pp. 931–940, 2013.

[16] G. Pilia, W.-M. Chen, A. Scuteri et al., "Heritability of cardiovascular and personality traits in 6,148 Sardinians," *PLoS Genetics*, vol. 2, no. 8, article e132, 2006.

[17] J. van Dongen, G. Willemsen, W.-M. Chen, E. J. C. de Geus, and D. I. Boomsma, "Heritability of metabolic syndrome traits in a large population-based sample," *Journal of Lipid Research*, vol. 54, no. 10, pp. 2914–2923, 2013.

[18] F. V. Rijsdijk and P. C. Sham, "Analytic approaches to twin data using structural equation models," *Briefings in Bioinformatics*, vol. 3, no. 2, pp. 119–133, 2002.

[19] J. Lim, J. Sung, and S. Won, "Efficient strategy for the genetic analysis of related samples with a linear mixed model," *Journal of the Korean Data and Information Science Society*, vol. 25, no. 5, pp. 1025–1038, 2014.

[20] A. R. Gilmour, R. Thompson, and B. R. Cullis, "Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models," *Biometrics*, vol. 51, no. 4, pp. 1440–1450, 1995.

[21] D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding, "Improved heritability estimation from genome-wide SNPs," *The American Journal of Human Genetics*, vol. 91, no. 6, pp. 1011–1021, 2012.

[22] T. Thornton, H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan, and N. Risch, "Estimating kinship in admixed populations," *The American Journal of Human Genetics*, vol. 91, no. 1, pp. 122–138, 2012.

[23] C. R. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, John Murry, London, UK, 1861.

[24] I. C. McManus and C. G. N. Mascie-Taylor, "Human assortative mating for height: non-linearity and heteroscedasticity," *Human Biology*, vol. 56, no. 4, pp. 617–623, 1984.