이학석사 학위논문

Joint Selection of SNPs for Improving Prediction in Genome-wide Association Studies

전장유전체 연관분석에서의 예측력 향상을 위한 단일염기다형성 지표의 복합적 발굴 방법론

2013 년 2 월

서울대학교 대학원

통계학과

방 서 진

Joint Selection of SNPs for Improving Prediction in Genome-wide Association Studies

지도 교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함
2013 년 2 월

서울대학교 대학원
통계학과
방 서 진

방서진의 이학석사 학위논문을 인준함
2013 년 2 월

위 원 장 전 종 우 (인)

부위원장 박 태 성 (인)

위 원 조 신 섭 (인)

Abstract

Joint Selection of SNPs for Improving Prediction in Genome-wide Association Studies

Seo-Jin Bang

Department of Statistics

The Graduate School

Seoul National University

It is of great interest to select single-nucleotide polymorphisms (SNPs) associated with diseases in genome-wide association studies (GWAS). Since genetic variants affect diseases in multiple ways, the joint analysis of SNPs is important to understand the full effects of genetic variants. However, since the number of SNPs is large and there exists linkage disequilibrium (LD) among SNPs, it is not easy to identify the joint effects of SNPs on diseases. Thus, the multi-step approach is commonly used for handling these problems:

the ‘large p and small n’ problem and LD structure among SNPs. First, SNPs marginally associated with diseases are selected via single SNP analysis. Next, joint identification of putative SNPs via penalized regularization method is carried out for the pre-selected SNP set. Finally, SNPs from the joint identification step are ordered by a measure which is yielded from the joint analysis. Some current approaches have proposed scoring measures to select causal SNPs such as selection stabilities and effect sizes. In this dissertation, we discuss some pros and cons of the scoring measures and propose new joint SNP selection measures based on re-sampling methods such as permutation and bootstrap. We illustrate the joint SNP selection based on our measure by using bipolar disorder data from Welcome Trust Case Control Consortium (WTCCC). We demonstrate that the proposed method substantially improves the prediction of disease status compared to other scoring measures.

Keywords : Genome-wide association study (GWAS); Welcome trust case control consotrium (WTCCC); bipolar disease; joint selection via elastic net; permuted p-value

Student Number : 2011 - 20245

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

Together with methodological developments and advances in genotyping technologies, genome-wide association studies (GWAS) have provided advantages for detection of genetic variants associated with common human diseases [1, 2, 3]. In several years, more than 1600 genome wide associations have been published for 240 traits [4].

However, a large amount of GWAS data brings new difficulties in data storage, manipulation and analysis. Other difficulties for GWAS arise because of multiple loci mapping. Since genetic variants affect diseases in multiple ways, selecting SNPs based on marginal effect of a marker might bring quite different results from their true effects [2, 5, 6, 7]. For example, SNPs which are not actually associated with disease but are highly correlated to causal SNPs would increase false discovery rate. Also there exist actual causal SNPs which are not marginally correlated with disease but are jointly associated with disease.

It is computationally infeasible to perform joint analysis of all SNPs at once, because the number of SNPs in GWAS is extremely large enough to overwhelm the number of samples.

One efficient strategy for handling the 'large p and small n' problem is the use of the multi-step approaches [6, 7, 8]. At the first step, which is referred as a prescreening step, SNPs which are marginally associated with disease are selected to compose the candidate SNP set. This step helps to reduce dimensionality from extremely high to moderate scale that is small enough to carry out joint analysis. Most commonly used methods in prescreening step are sure independence screening (SIS) and iterative sure independence screening (ISIS) procedure [9]. These methods were shown to speed up variable selection process while keeping all the important SNPs [9]. Next, joint identification is performed on the moderate number of candidate SNPs which survive after the prescreening step. However, identifying SNPs in high dimensional setting and handling linkage disequilibrium (LD) among SNPs still remained as difficult challenges. Fortunately, several penalized methods have successfully addressed the LD problems including bridge, the least absolute shrinkage and selection operator (LASSO), ridge, smoothly clipped absolute deviation (SCAD), adaptive lasso, and elastic-net (EN) [10 - 14]. In order to sort out meaningful SNPs, SNPs which are selected at the joint analysis step are ordered and reselected by a measure yielded from the joint analysis step.

Since the distribution of penalized estimators has not been derived well, alternative approaches are needed to access the

significance of each SNP. Recently, some studies have proposed measures based on the concept of selection consistency. Selection consistency assumes that the selected SNP set contains causal SNPs with the probability of 1 when the variable selection via penalized method is repeatedly applied to independent training samples. Necessary and sufficient conditions for selection consistency of several penalized methods, such as LASSO, adaptive LASSO, SCAD, and EN, to consistently select causal SNPs have been investigated [12, 13, 14, 16].

As a measure of selection consistency, selection stability is commonly used which is defined as a proportion of how many times each identified SNP is replicated via an automatic selection procedure such as penalized regularization in independent training samples [7, 17]

However, generating enough number of independent training sets is impracticable under a limited sample size. Resampling procedures such as bootstrap and permutation are commonly used to generate training sets [6, 7, 18].

While these resampling procedures are simple and can be applied to the case with a limited size of samples, for estimating the selection stability, they fail to obtain reliable measures for identifying causal SNPs. The resampled training sets do not satisfy independency which should be guaranteed for the selection stability. Moreover, the existences of selection

consistency of penalized models are assured only if the size constraints of the shrinkage parameter are satisfied [12, 13, 14, 16].

Another frequently used approach is scoring SNPs by the effect size of a bootstrap coefficient estimator which is defined as the mean of coefficient estimators yielded from bootstrap samples. However, when a penalized method is used to estimate the effect size of a bootstrap coefficient, the bias for true effect size of SNPs arises [19].

1.2 Overview

In this dissertation, we present new scoring measures based on the multi-step approach. At the first stage, the top k SNPs which are marginally associated with disease are pre-selected. This prescreening helps to reduce ultrahigh dimensionality and improve computational efficiency. Next, automatic variable selection via penalized regression method is performed for the top k SNPs. We use EN regularization method for variable selection. EN selects SNPs automatically and handles a severe multicollinearity problem [10]. EN regularization method is also applied to the permuted and bootstrapped data sets. Several new scoring measures are then proposed by using permuted p-value, bootstrap selection stability, bootstrap effect size and adjusted bootstrap effect size

calculated from these resampled data sets. These scoring measures are used to sort out meaningful SNPs.

In order to evaluate these scoring measures, real data analysis using Welcome Trust Case Control Consortium (WTCCC) bipolar disease (BD) data is implemented. Several prediction methods such as Linear Discriminant Analysis (LDA), EN, Random Forest (RF), and Support Vector Machine (SVM) are used. We focus on the area under the curves (AUC) to compare prediction performance of the various scoring measures.

Chapter 2. Materials and Methods

2.1 Prescreening

Single SNP analysis using the logistic regression model (1) is implemented for all SNPs. When there is population stratification in the data, adjusting variables should be added to the model.

$$\text{logit}(\mu_i) = \gamma_0 + \sum_{q=1}^{Q} \gamma_q Z_q + \beta_j SNP_{ij} \qquad (1)$$

Note that Z_q , SNP_{ij} and μ_i represent the $q-$th adjusting variable, the number of minor alleles in the $j-$th SNP and expectation of binary trait, respectively; $i = 1, 2, \cdots, n$ and n is the number of samples; $j = 1, 2, \cdots, J$ and J is the number of all SNPs; $q = 1, 2, \cdots, Q$ and Q is the number of adjusting variables. All SNPs are ranked in ascending order of the p−value. The top k SNPs showing strongest marginal associations with the trait are selected. The prescreening releases a computational burden which is induced by an extremely large number of SNPs, and enables us to perform the joint analysis on the reduced SNP set.

2.2 Joint selection of SNPs via Elastic-Net

A multiple logistic model (2) is fitted to the top k SNPs which are selected at the prescreening step.

$$logit(\mu_i) = \gamma_0 + \sum_{q=1}^{Q} \gamma_q Z_q + \sum_{j=1}^{k} \beta_j SNP_{ij} \quad (2)$$

EN regularization method is used for estimating optimal solution. EN finds $\hat{\beta}$ which minimizes $-L(\beta, \gamma) + \lambda P_\alpha(\beta)$ subject to α and λ constraints. The EN penalty is defined as (3).

$$P_\alpha(\beta) = \frac{1}{2}(1-\alpha) \| \beta \|_{l2}^2 + \alpha \| \beta \|_{l1} \quad (3)$$

Note that α controls the weight on the penalty of lasso in contrast to the penalty of ridge. EN penalized regression shrinks the regression coefficients toward the origin by imposing the penalty (3). Parameter λ represents the amount of shrinkage which is implemented by the EN penalty. Ten-fold cross-validation is used to find an optimal λ which maximizes the mean squared error. The EN penalty (3) is a convex combination of lasso [11] and ridge [12] penalties. Therefore, EN takes advantages of both ridge and lasso regularization methods. Automatic dimension reduction is performed by exploring sparse solution, and provides robust coefficient estimates under the existence of extreme correlations among

SNPs. Highly correlated SNPs tend to be selected together [10].

2.3 Empirical replication of joint analysis on resampled data sets

Training data sets are generated using resampling procedures such as bootstrap and permutation. The bootstrap data sets are generated by sampling individuals with replacement from the original sample. They can be regarded replicated samples from the population itself [21]. The bootstrap data set keeps a biologically meaningful association which the original data set has.

The permuted data sets are generated by shuffling labels of the binary trait. EN joint analysis is performed on these resampled data sets. Shuffling makes data sets to have no biologically meaningful association with the trait [2]. The distribution of the coefficient estimators from permuted data sets can be used as an approximation to the null distribution of the coefficient estimator. Since the null distributions of penalized estimators have not been well derived, an empirical distribution based on the permuted data sets is a useful approximation to a null distribution of a penalized estimator.

2.4 A scoring measure based on permuted samples

In order to rank SNPs which are survived at the joint selection step via EN, we propose the permuted p−value as a scoring measure. Note that the permuted p−value is calculated for each selected SNP. The permuted p−value is defined as the following.

$$\mathrm{PMP_j} = \frac{1}{P}\sum_{p=1}^{P} I(|\widehat{\beta_j^{\mathrm{p}}} - \overline{\beta_j^{\mathrm{p}}}| > |\hat{\beta}_j - \overline{\beta_j^{\mathrm{p}}}|) \qquad (4)$$

$\hat{\beta}_j$, $\widehat{\beta_j^{\mathrm{p}}}$ and $\overline{\beta_j^{\mathrm{p}}}$ represent an estimator of j−th SNP in the original data set, an estimator of j−th SNP in the p−th permuted data set and the mean of $\widehat{\beta_j^{\mathrm{p}}}$, respectively; P is the number of permuted data sets; $I(\cdot)$ is an indicator function. When a SNP is not selected at the p−th permuted set, the coefficient of the SNP is considered to be estimated as zero. For every SNP, there exists P number of coefficient estimates including the zero estimates. A SNP with a small number of zero estimates tends to have a small PMP.

Note that penalized coefficient estimator is a biased estimator of the coefficient, which means the expected value of penalized estimator can be expressed as (5)

$$\mathrm{E}\left(\widehat{\beta_j^p}\right) = \beta_j + bias(\widehat{\beta_j^p}) \quad (5)$$

β_j and $bias(\widehat{\beta_j^p})$ represent the true effect size of the j−th SNP and bias of $\widehat{\beta_j^p}$, respectively. Therefore, the null hypothesis $H_0 : \beta_j = 0$, which means no causal effect of the j−th SNP on the disease status, is equivalent to $\mathrm{H_0} : \mathrm{E}\left(\widehat{\beta_j^p}\right) = bias\left(\widehat{\beta_j^p}\right)$. Therefore, a test statistic should be defined by adjusting the bias. As shown in (4), PMP adjusts bias by subtracting $\overline{\beta_j^p}$ which is used as a substitute for $\mathrm{E}\left(\widehat{\beta_j^p}\right)$ from each side.

2.5 Scoring measures based on bootstrap samples

Another resampling procedure, bootstrap, is used to sort out meaningful SNPs from the joint selection step via EN. We first summarize several scoring measures based on the bootstrap data sets.

2.5.1 Bootstrap selection stability (BSS)

BSS [7] uses a bootstrap resampling procedure to generate

training sets which are used to estimate selection consistency of a SNP. Bootstrap data sets with the same size of the original data set are generated and joint selection of SNPs via EN is performed for each bootstrap set. BSS represents the frequency how many times each identified SNP is replicated via penalized method in the bootstrap sets.

$$BSS_j = \frac{1}{B}\sum_{b=1}^{B} I_j^b \quad (6)$$

Note that I_j^b is a function indicating whether the j−th SNP is selected through joint selection via EN on the b−th bootstrap data set or not. B represents the number of bootstrap data sets. SNPs are ranked in descending order of BSS.

2.5.2 Bootstrap effect size (BES)

BES [19] is obtained by averaging estimated effect sizes of the bootstrap coefficient estimator. BES also utilizes the result of joint analysis on the bootstrap sets which are used to calculate BSS.

$$BES_j = \frac{1}{B}\sum_{b=1}^{B} \widehat{\beta_j^b} \quad (7)$$

Note that $\widehat{\beta_j^b}$ is the estimator of j−th SNP at the b−th bootstrap samples. If a SNP is not selected at the b−th

bootstrap data set, its coefficient is considered to be zero at the b−th bootstrap data set. SNPs are ranked in descending order of the absolute value of BES.

2.5.3 Adjusted BES (adjBES)

For adjusting for the average effect size, adjBES divides BES by its estimated standard deviation.

$$\text{adjB}ES_j = \frac{BES_j}{\sqrt{\widehat{Var}(\widehat{\beta_j^b})/B}} \quad (8)$$

where $\widehat{Var}\left(\widehat{\beta_j^b}\right) = \sum_{b=1}^{B}\left(\widehat{\beta_j^b} - BES_j\right)^2/B-1$. Note that the unselected SNP is considered to have a zero estimate. SNPs are ranked in descending order of the absolute value of adjBES.

Chapter 3. Results

The multi-step approach using new scoring measures is applied to a WTCCC dataset with 4,806 individuals and 354,022 SNPs which passed the quality control process. Top k SNPs which are ordered by the scoring measures are used for predicting disease status. Several prediction methods such as LDA, EN, RF, and SVM are used to predict bipolar disorder (BD) status. The AUC of various scoring measures are compared with each other.

3.1 Data

WTCCC [1] is a consortium for genome-wide association studies of 14,000 cases and 3,000 shared controls in the British population. Each 2,000 of 14,000 cases are composed of 7 complex human diseases-BD, coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). In this analysis we focused on BD.

BD, a manic depressive illness [20], is a dangerous disease usually accompanied by disturbances in thinking and behavior, and may lead to serious injury or even death. Sibling

recurrence risk (λs) ranges from 7 to 10 and heritability does from 0.8 to 0.9 [21, 22], which supports that inherited genetic variation plays a major role in causing BD. There have been several genetics studies investigating the effect of multiple risk alleles on BD [23, 24].

3.2 Quality Control

Samples were genotyped on the Affymetrix GeneChip 500K Mapping Array Set [1]. Quality Control process is implemented for 5000 individuals (2000 BD 3000 Control) with 500,568SNPs. SNPs with minor allele frequency (MAF) < 5%, missing genotype frequency > 5%, hardy weinberg equilibrium (HWE) in all 2 control cohorts < $5.7*10^{-7}$, allelic and/or genotypic association between 2 control cohorts p−value < $5.7*10^{-7}$ are excluded. Imputation for missing genotype is implemented using fastPHASE. After the quality control process, 4,806 individuals (1,868 cases of BD and 2,938 controls) and 354,022 SNPs are included in the analysis.

3.3 Analysis of WTCCC Data

At the prescreening step, single SNP analysis is performed for 354,022 SNPs which passed the quality control process. Population stratification refers to differences in allele

frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease [2]. In order to control false discovery rate resulted from the population stratification, we include four adjusting covariates –sex, age, and two principle components– to the model (9).

$$\mathrm{logit}(\mu_i) = \gamma_0 + \gamma_1 SEX + \gamma_2 AGE + \gamma_3 PC1 + \gamma_4 PC2 + \beta_j SNP_{ij} \quad (9)$$

Note that $i=1, 2, \cdots, 4806$ and $4,806$ is the number of individuals; $j=1, 2, \cdots, 354022$ and $354,022$ is the number of SNPs. Since that BD is a binary trait, a logit function is used as a link function. A thousand of SNPs were pre-selected in the ascending order of p-values from single SNP analysis. The result of the prescreening is shown in Fig.1: minus-log transformed p-values for 354,022 SNPs are dotted. The x-axis represents geographical region of SNP in each chromosome. SNPs above the horizontal line are the top 1,000, 2,000, 3,000, and 4,000 SNPs from top to bottom which are used to compose the candidate SNP set for joint analysis step (Fig. 1)

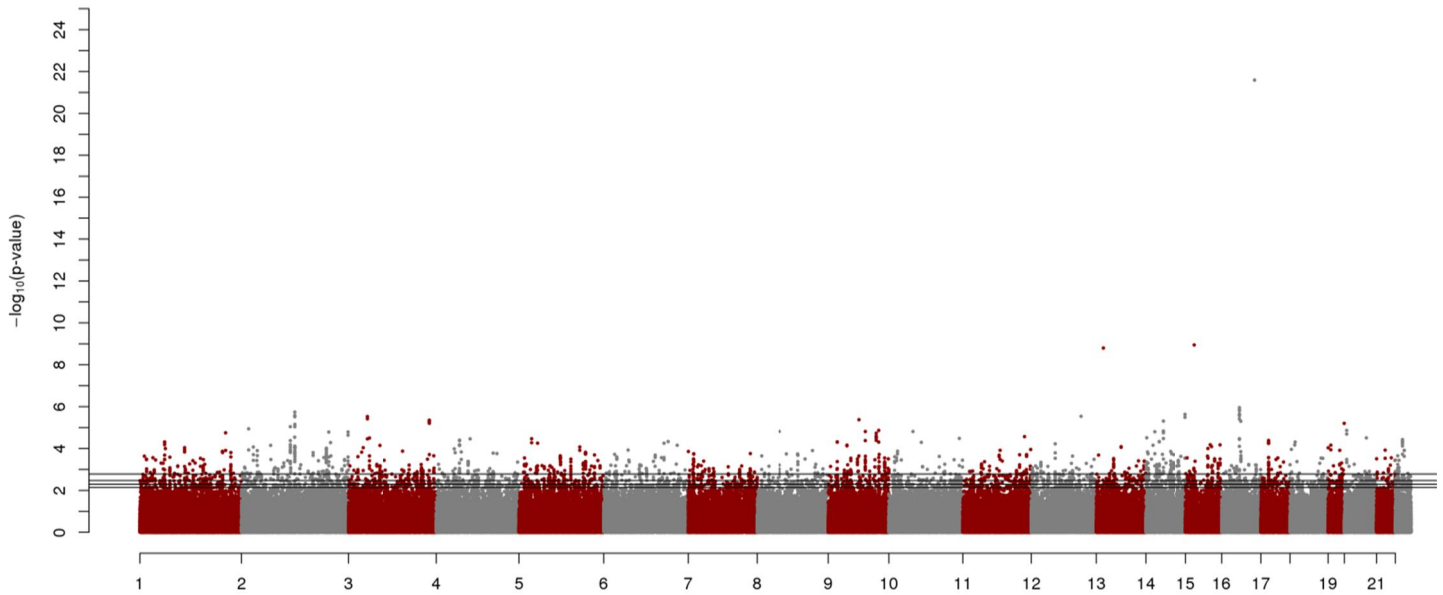

Figure 1. P-values from single SNP analysis using logistic model are dotted on the Manhattan plot. The x-axis represents geographical region of SNP in each chromosome. SNPs above the horizontal line are the top 1000, 2000, 3000, and 4000 SNPs from top to bottom which are used for the Elastic-net analysis after prescreening.

For joint selection of SNPs, a multiple logistic model was fitted for top 1,000, 2,000, 3,000 and 4,000 SNPs. EN regularization method was used to estimate coefficients.

$$logit(\mu_i) = \gamma_0 + \gamma_1 SEX + \gamma_2 AGE + \gamma_3 PC1 + \gamma_4 PC2 + \sum_{j=1}^{k} \beta_j SNP_{ij} \quad (10)$$

For each k number of prescreened SNPs, the value λ was chosen by ten-fold cross-validation, which minimized the mean squared error (Table 1). We assumed a tuning parameter $\alpha = 0.5$. After joint selection via EN, SNPs showing strong joint association with the BD status were automatically selected.

The number of Prescreened SNPs	λ	The number of Selected SNPs
1000	0.00365	639
2000	0.00309	1159
3000	0.00252	1560
4000	0.00109	1988

Table 1. The optimal tuning parameter λ, and the number of remained SNPs after the joint selection of SNPs according to the number of prescreened SNPs.

A thousand of permutation samples with size 4,806 were generated by shuffling the labels of phenotype. Then, EN

regularization method was then applied to the 1,000 permuted data sets and 1,000 bootstrapped data sets. We used the same fixed values of λ presented at Table 1. Then, the scoring measures were computed such as PMP from permutation data sets and BSS, BES, and adjBES from bootstrap data sets. The SNPs, survived from both the prescreening and the joint analysis via EN, were ordered by the scoring measures.

In order to compare the scoring measures, we construct prediction models using the same number of SNPs which are ordered by different measures. Other scoring measures such as p-value from single SNP analysis and effect size of coefficient estimator are also compared to the scoring measures based on resampling procedures. Since EN penalized method provides automatic variable selection as a lasso penalized method does, we select the same number of SNPs by controlling the tuning parameter λ. Different values of the tuning parameter λ are used to select the different numbers of SNPs instead of the fixed value which is presented at Table 1. Several prediction methods such as LDA, EN, RF, and SVM are used to predict BD status. AUC are calculated by tenfold cross-validation.

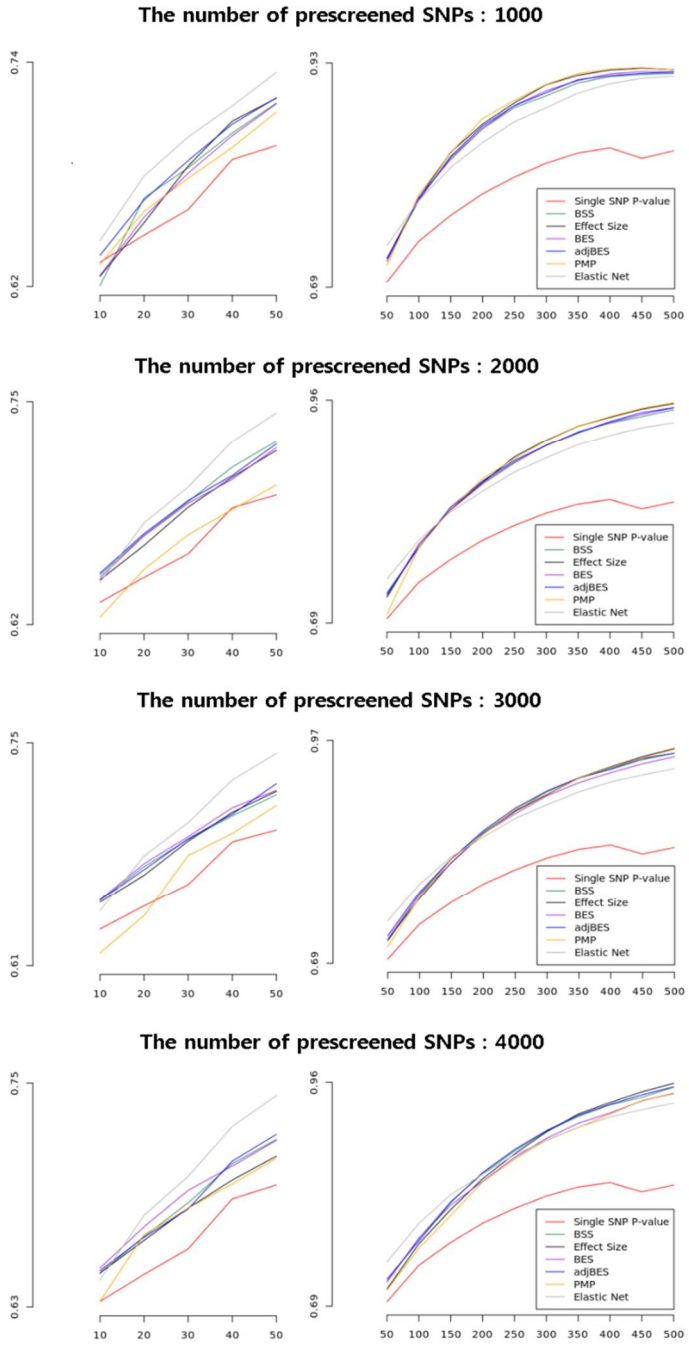

Figure 2. AUC via LDA. The x-axis represents the number of SNPs which are used in the prediction models. Each measures are identified by colors.

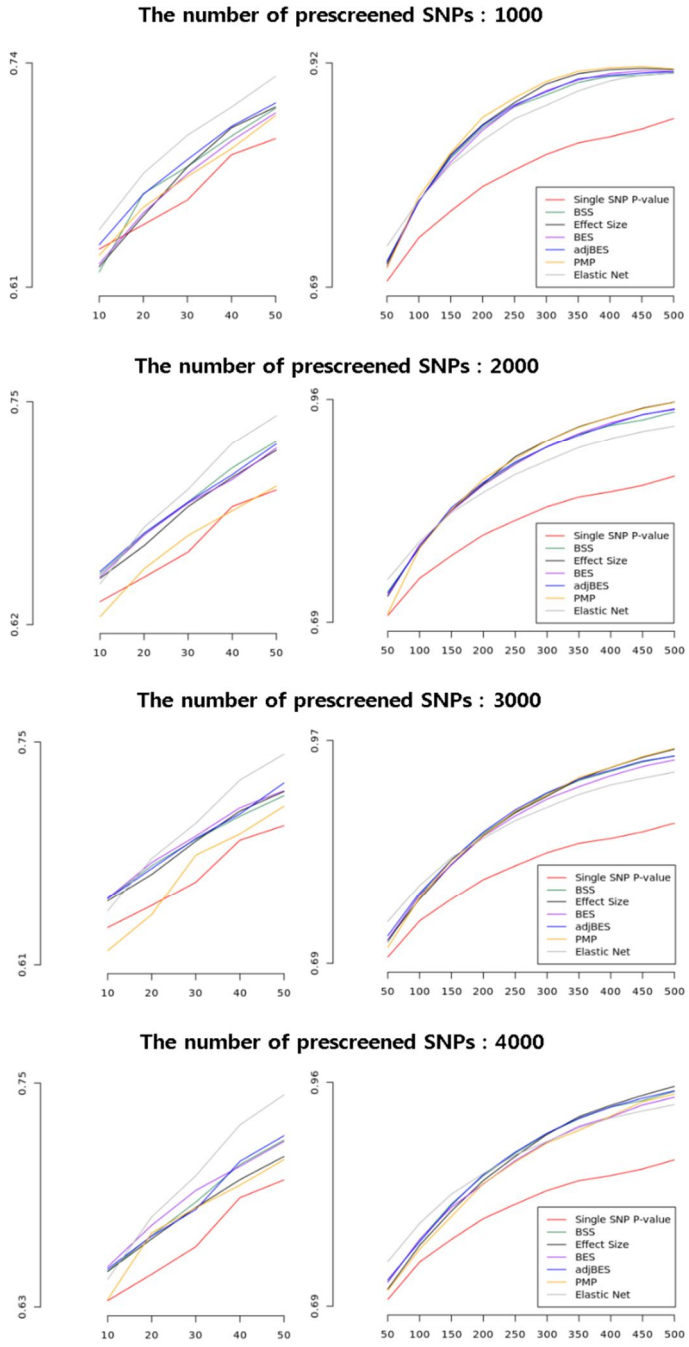

Figure 3. AUC via EN. The x-axis represents the number of SNPs which are used in the prediction models. Each measures are identified by colors.

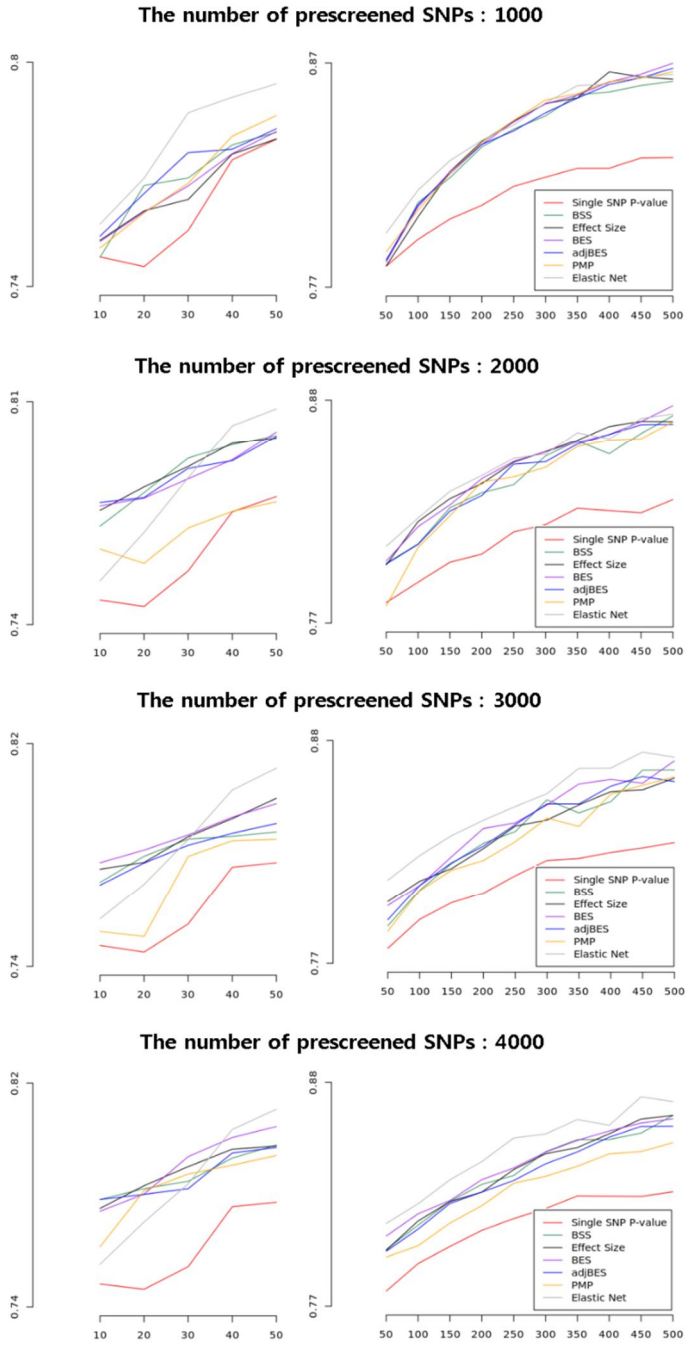

Figure 4. AUC via RF. The x-axis represents the number of SNPs which are used in the prediction models. Each measures are identified by colors.

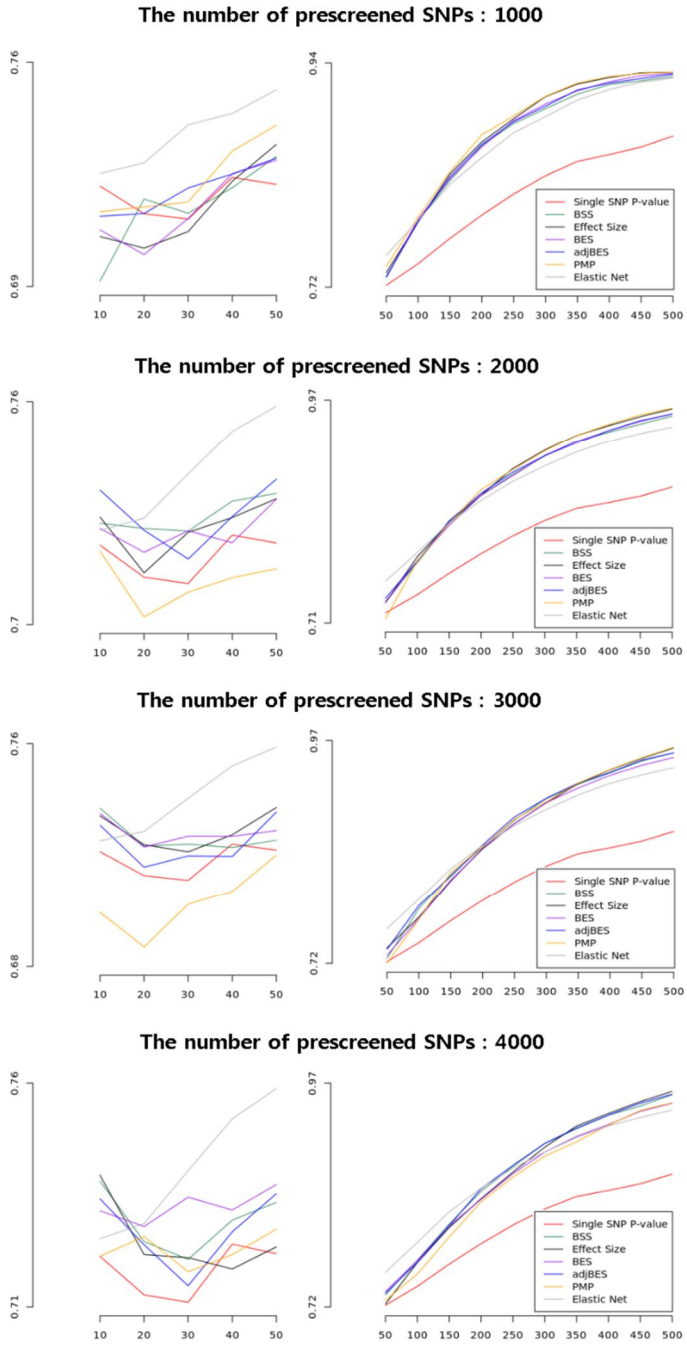

Figure 5. AUC via SVM. The x-axis represents the number of SNPs which are used in the prediction models. Each measures are identified by colors.

The right sides of Fig. 2 – Fig. 5 plot the AUC of each scoring measure from 50 to 500 SNPs with 50 intervals. The left sides of Fig. 2 – Fig. 5 plot the AUC of each scoring measure from 10 to 50 SNPs with 10 intervals. The AUC increases as the number of SNPs increases, which is shown in both sides of Fig.2–Fig.5. In the prediction models using more than 50 SNPs, scoring measures based on joint analysis tend to perform overwhelmingly better than that based on single SNP analysis. However, when the number of SNPs is relatively small, the ranking fluctuation makes it difficult to identify a ‘leading measure’ showing the highest AUC.

The AUC of scoring measures calculated from RF shows considerably different trends from those calculated from other prediction methods such as LDA, EN, and SVM. While the rank fluctuates when the small number of SNPs is used to construct the LDA, EN, and SVM prediction models, PMP dominates the other scoring measures when relatively large numbers of SNPs are used to construct these prediction models. However, in the RF, PMP shows similar or even lower AUC than other scoring measures which are based on joint analysis. As seen in the right sides of Fig. 2 - 5, the AUC of PMP decrease as the number of prescreened SNPs increases in all prediction models.

In Fig. 2, Fig. 3, and Fig. 5 using more than 50 SNPs, scoring measures, such as BSS, BES, and adjBES, which are based on a bootstrap resampling procedure, show similar or

even lower AUC than scoring measures based on the effect size. Note that the effect size can be easily yielded from the joint selection on the original data. It means that the scoring measures based on bootstrap samples are not useful in terms of both computational efficiency and prediction performance in the LDA, EN, and SVM prediction models.

The AUC of prediction model consisting of SNPs, which are selected via the EN automatic selection, is labeled as *Elastic Net*. In LDA, EN, and SVM using relatively small number of SNPs, the Elastic Net automatic selection shows higher AUC than other scoring measures. However, as the number of SNPs in these predictions models increases, the Elastic Net automatic selection shows lower AUC than other scoring measures. Since EN regularization method finds coefficient estimates which minimize $-L(\beta,\gamma)+\lambda P_{\alpha}(\beta)$, the λ is the weight on the elastic−net penalty in contrast to $-L(\beta,\gamma)$. As the weight λ decreases, the number of automatically selected SNPs increases. Therefore, as the weight on the elastic−net penalty decreases, the AUC of the Elastic Net automatic selection, which is calculated from LDA, EN, and SVM, becomes less than the AUC of scoring measures based on the resampling procedure.

However, in the RF using more than 50 SNPs, as the number of prescreened SNPs increases, the Elastic Net scoring measure shows higher values than other scoring measures based on resampling procedures.

The Elastic Net automatic selection is fast in selecting the fixed number of SNPs. When the number of selected SNPs is pre-determined, the Elastic Net automatic selection spends less time than other scoring measures which are based on the resampling procedure. The Elastic Net automatic selection does not spend time on implementing EN on the resampled data. The Elastic Net automatic selection only spends time on searching an appropriate lambda to select the number of pre-determined SNPs, which is similar to explore optimal lambda for other scoring measures such as BSS, BES, adjBES, and PMP.

In the LDA, EN, and SVM, the Elastic Net is more efficient in the computational and predictive manner when the weight λ is larger than the smallest λ in Tables 2, 4, and 6. However, as seen in Tables 3, 5, and 7, the largest number of selected SNPs for which the Elastic Net automatic selection is more efficient than other scoring measures, depends on the number of prescreened SNPs. For example, in Table 3, the Elastic net automatic selection requires 3,000 numbers of prescreened SNPs to select 150 SNPs satisfying that the AUC is larger than the AUC of other scoring measures in LDA.

On the other hand, scoring measures, which are based on the resampling procedures, have larger AUC when the weight λ is smaller than the smallest λ in Tables 2, 4, and 6. Moreover, these scoring measures can rank all selected SNPs, while the Elastic Net cannot.

	top 1000	top 2000	top 3000	top 4000
BSS	0.05395	0.05077	0.04863	0.04865
BES	0.05395	0.05077	0.04863	0.04427
adjBES	0.05395	0.05077	0.04863	0.04865
PMP	0.05395	0.05077	0.04863	0.04160

Table 2. The smallest λ satisfying that the AUC of the Elastic Net is larger than the AUC of other scoring measures in LDA.

	top 1000	top 2000	top 3000	top 4000
BSS	50	100	150	150
BES	50	100	150	250
adjBES	50	100	150	150
PMP	50	100	150	350

Table 3. The largest number of selected SNPs satisfying that the AUC of the Elastic Net is larger than the AUC of other scoring measures in LDA

	top 1000	top 2000	top 3000	top 4000
BSS	0.05067	0.05077	0.04863	0.04603
BES	0.05067	0.05077	0.04863	0.04273
adjBES	0.05067	0.05077	0.04863	0.04603
PMP	0.05395	0.05077	0.04863	0.04160

Table 4. The smallest λ satisfying that the AUC of the Elastic Net is larger than the AUC of other scoring measures in EN.

	top 1000	top 2000	top 3000	top 4000
BSS	100	100	150	200
BES	100	100	150	300
adjBES	100	100	150	200
PMP	50	100	150	350

Table 5. The largest number of selected SNPs satisfying that the AUC of the Elastic Net is larger than the AUC of other scoring measures in EN

	top 1000	top 2000	top 3000	top 4000
BSS	0.05077	0.05077	0.04602	0.04603
BES	0.05395	0.04849	0.04602	0.04273
adjBES	0.05077	0.05077	0.04863	0.4603
PMP	0.05395	0.04849	0.04602	0.04160

Table 6. The smallest λ satisfying that the AUC of the Elastic Net is larger than the AUC of other scoring measures in SVM.

	top 1000	top 2000	top 3000	top 4000
BSS	100	100	200	200
BES	50	150	200	300
adjBES	100	100	150	200
PMP	50	150	200	350

Table 7. The largest number of selected SNPs satisfying that the AUC of the Elastic Net is larger than the AUC of other scoring measures in SVM

Chapter 4. Discussion

In this dissertation, we present new scoring measures based on the multi-step approach. At the first stage, the top k SNPs which are marginally associated with disease are pre-selected. Next, automatic variable selection via penalized regression method is performed for the top k SNPs. EN regularization method is also applied to the permuted and bootstrapped data sets. Several new scoring measures are then proposed by using permuted p-value, bootstrap selection stability, bootstrap effect size and adjusted bootstrap effect size calculated from these resampled data sets. We discuss some pros and cons of the scoring measures and propose new joint SNP selection measures based on re-sampling methods such as permutation and bootstrap.

There is always a trade-off between computational burden and estimation accuracy in selecting causal SNPs on high-dimensional data. In order to handle these problems we implement a prescreening step. It is based on the sure independence screening (SIS) method [9]. In order to release computational burden while keeping all the important SNPs, the size of prescreened SNP set should be small enough to make computation feasible and big enough to contain susceptible SNPs. It has been shown that all the important variables survive

after applying SIS method. Fan and Lv [9] presented $n-1$ or $n \times \log(n)$ as the number of prescreened SNPs. For determining the optimal number of pre-selected SNPs, we performed the multi-step approach for the different numbers of pre-selected SNPs which ranges from 1000 to 4000.

There are several advantages of using EN penalty for joint SNP selection. The EN penalty is combination of ridge and lasso, thus it takes advantages of both ridge and lasso penalties [10]. The lasso term plays a part in automatic SNP selection especially when there exists only a relatively small number of causal SNPs [11]. However, the existence of multicollinearity among SNPs cannot be revealed by the lasso term alone. The ridge term plays a part in dealing with this problem and provides stable estimation. EN is known to outperform lasso in terms of selection consistency [10, 11, 14, 16]. Moreover, it has been shown that EN consistently selects the true model even under the sparsity condition, where the total number of predictors and the sample size go to infinity [25]. Any penalization method selecting SNPs automatically can be used instead of EN at the joint analysis step. Scoring measures are calculated from the joint analysis step via the penalized method.

We proposed several scoring measures for a joint SNP selection based on resampling procedures such as permutation and bootstrap methods. The difference between bootstrap and permutation methods is whether or not the association with

disease is the same as the original data. Bootstrap data keep biologically meaningful association which original data set has. Thus, the scoring measures based on the bootstrap data such as BSS, BES, and adjBES can be considered to be derived from the distribution under the alternative hypothesis. On the other hand, when labels of case and control are shuffled, the biological association disappears [2]. Thus, the permuted data can be used for estimating the null distributions of coefficients. Especially it is useful for estimating the p-values when the probability distribution of the coefficient is difficult to handle. Since most distributions of penalized estimators are not easy to derive, the empirical distribution based on permuted data sets can be useful to estimate the null distributions of penalized estimators. PMP can be used as an empirical estimator of p-value of penalized coefficients. As shown in (4), bias of penalized estimators is adjusted by subtracting the permutation sample mean $\overline{\beta_j^{\mathrm{p}}}$ from $\hat{\beta}_j$ and $\widehat{\beta_j^{\mathrm{p}}}$.

Even though scoring measures which are based on resampling procedures are less efficient in computational manner than the Elastic Net automatic selection, these measures are recommended to select relatively large numbers of SNPs, which depends on the number of prescreened SNPs. The Elastic Net automatic selection is computationally more efficient than the scoring measures. However its prediction performance depends on the number of prescreened SNPs. The

Elastic Net automatic selection has a larger AUC only when the weight on EN penalty is relatively large.

For evaluation of the scoring measures, real data analysis was performed on BD phenotype of WTCCC and prediction performances were compared to each other. Our analysis suggests that PMP tends to have better prediction accuracy and AUC than other scoring measures when using relatively large numbers of SNPs in LDA, EN, and SVM. However, on the prediction models using a relatively small number of SNPs, the leading measure which is continually showing the highest prediction accuracy regardless of the number of SNPs does not exist.

The recommended number of resampled data sets depends on the number of SNPs which are selected from the joint analysis. When the number of resampling is small, the resampled data set tends to assign a tied score to different SNPs. Even when the number of resampling is larger than the number of SNPs, there might exist SNPs which have a tied value of scoring measure. We handle this tied problem by ranking the tied SNPs in descending order of the effect size.

Bibliography

[1] The Wellcome Trust Case Control Consortium, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls,” Nature, vol. 447, Jun. 2007, pp. 661-678, doi:10.1038/nature05911.

[2] Joel N.Hirschhorn and Mark J.Daly, “Genome-wide association studies for common diseases and complex traits.” Nat. Rev. Genet., vol. 6, Feb. 2005, pp 95-108, doi:10.1038/nrg1521.

[3] William Y. S.Wang, Bryan J. Barratt, David G. Clayton and John A. Todd, “Genome-wide association studies: theoretical and practical concerns,” Nat. Rev. Genet., vol. 6, Feb. 2005, pp 109-118, doi:10.1038/nrg1522.

[4] http://www.genome.gov/gwastudies.

[5] Josephine Hoh and Jurg Ott , “Mathematical multi-locus approaches to localizing complex human trait genes,” Nat. Rev. Genet., vol. 4, Sep. 2003, pp 701-709, doi:10.1038/nrg1155.

[6] Qianchuan He and Dan-Yu Lin, “A variable selection method for genome-wide association studies,” Bioinformatics, vol. 27, Jan. 2011, pp 1–8, doi:10.1093/btq600.

[7] Seoae Cho, Kyunga Kim, Young Jin Kim, Jong-Keuk Lee, Yoon Shin Cho, Jong-Young Lee, Bok-Ghee Han, Heebal

Kim, Jurg Ott and Taesung Park, “Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis,” Ann. of Hum Genet., vol. 74, Sep. 2010, pp. 416–428, doi: 10.1111/j.1469-1809.2010.00597.x.

[8] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel and Kenneth Lange, “Genome-wide association analysis by lasso penalized logistic regression,” Bioinformatics, vol. 25, Mar. 2009, pp 714–721, doi:10.1093/bioinformatics/btp041.

[9] Jianqing Fan and Jinchi Lv, “Sure Independence Screening for Ultra-High Dimensional Feature Space,” J. R. Stat. Soc. Series B Stat. Methodol., vol. 70, Oct. 2008, pp 849–911, doi:10.1111/j.1467-9868.2008.00674.x.

[10] Hui Zou and and Trevor Hastie, “Regularization and variable selection via the elastic net,” J. R. Statist. Soc., vol. 67, Apr. 2005, pp. 301–320, doi: 10.1111/j.1467-9868.2005.00503.x.

[11] Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso,” J. R. Statist. Soc., vol. 58, 1996, pp 267–288, doi:10.2307/2346178.

[12] S. Le Cessi and J. C. Van Houwelingen, “Ridge Estimators in Logistic Regression,” Appl. Statist., vol. 41, 1992, pp 191-201, doi:10.2307/2347628.

[13] Jianqing Fan and Runze Li, “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,”

JASA, vol. 96, Dec. 2001, pp 1348–1360, doi:10.1198/016214501753382273.

[14] Hui Zou, “The adaptive lasso and its orachle properties,” JASA, vol. 101, 2006, pp 1418–1429, doi:10.1198/016214506000000735.

[15] Mee Young Park and Trevor Hastie, “Penalized logistic regression for detecting gene interactions,” Biostatistics, vol. 9, Jan. 2008, pp. 30–50, doi:10.1093/biostatistics/kxm010.

[16] Jinzhu Jia and Bin Yu, “On Model Selection Consistency of the Elastic Net When p ≫ n,” Statistica Sinica, vol. 20, 2010, pp 595–611.

[17] Benjamin A. Logsdon1, Cara L. Carty, Alexander P. Reiner, James Y. Dai, and Charles Kooperberg, “A novel variational Bayes multiple locus Z–statistic for genome–wide association studies with Bayesian model averaging,” Bioinformatics, vol. 28, Jul. 2012, pp 1738–1744, doi:10.1093/bioinformatics/bts261.

[18] Junhui Wang, "Consistent selection of tuning parameters in high–dimensional penalized regression,” unpublished.

[19] Suat Sahinler and Dervis Topuz, “Bootstrap and Jacknife Resampling Algorithms for Estimation of Regression Parameters,” Journal of Applied Quantitative Methods, vol. 2, Jun. 2007, pp 188–199.

[20] Bruno Müller–Oerlinghausen, Anne Berghöfer and Michael Bauer, “Bipolar disorder,” The Lancet, vol. 359, Jan.

2002, pp 241–247, doi:10.1016/S0140-6736(02)07450-0.

[21] N Craddock, M C O'Donovan, and M J Owen, "The genetics of schizophrenia and bipolar disorder:dissecting psychosis," J Med Genet, vol. 42, Jan. 2005, pp 193–204, doi:10.1136/jmg.2005.030718.

[22] Peter McGuffin, Fruhling Rijsdijk, Andrew, Pak Sham, Randy Katz and Alastair Cardno, "The heritability of bipolar affective disorder and the genetic relationship to unipolar depression," Arch. Gen. Psychiatry, vol. 60, May. 2003, pp 497-502, doi: 10.1001/archpsyc.60.5.497.

[23] V Moskvina, N Craddock, P Holmans, I Nikolov, J S Pahwa, E Green, Wellcome Trust Case Control Consortium, M J Owen, and M C O'Donovan, "Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk," Molecular Psychiatry, vol. 14, 2009, pp 252–260, doi:10.1038/mp.2008.133.

[24] Peter P. Zandi, Sebastian Zöllner, Dimitrios Avramopoulos, Virginia L. Willour, Yi Chen, Zhaohui S. Qin, Margit Burmeister, Kuangyi Miao, Shyam Gopalakrishnan, Richard McEachin, James B. Potash, J. Raymond DePaulo Jr., and Melvin G. McInnis, "Family-based SNP association study on 8q24 in bipolar disorder," American Journal of Medical Genetics Part B, vol. 147B, Jul. 2008, pp 612-618, doi: 10.1002/ajmg.b.30651.

[25] Tong Tong Wu and Kenneth Lange, “Coordinate descent algorithms for lasso penalized regression,” Ann. Appl. Stat., vol. 2, Mar. 2008, pp 224–244, doi:10.1214/07−AOAS147.

초 록

2003년 완료된 인간게놈프로젝트(human genome project: HGP)는 성공적으로 인간의 유전체 염기서열을 규명하고, 인간의 생명 현상을 결정짓는 유전자의 지도를 성공적으로 작성하였다. 이후 전장유전체를 대상으로 특정 질병과의 관련성을 밝히는 전장유전체 연관분석 (genome-wide association study: GWAS)이 폭발적으로 증가하였다. 특히 정상인과 복합질환이 있는 사람 간에 차이를 보이는 단일염기다형성 (single nucleotide polymorphism: SNP)을 이용한 질병지표 발굴에 대한 연구가 활발하게 이루어지고 있다.

SNP과 같은 유전적 변이는 특정 질병에 복합적으로 영향을 미치는 것으로 여겨진다. SNP이 가지는 영향력을 독립적으로 분석하는 것은 실제 생체 내에서 일어나는 복잡한 유전자 작용을 반영하지 못한다. 그러나 SNP 자료의 방대한 양을 처리함과 동시에 SNP 사이에 존재하는 연관 불균형 (linkage disequilibrium: LD)을 고려하는 것이 쉽지 않기 때문에 대부분의 연구에서 단일 SNP 분석 방법만을 사용하고 있다.

이와 같은 문제점을 해결하기 위해서 기존의 연구는 다음과 같은 다단계 접근 방법을 사용해 왔다. 첫 번째로 독립적인 SNP 분석을 통해 복합질환과 단일한 연관이 있는 SNP을 선별하여 고차원의 SNP 자료를 적절한 수준으로 차원 축소한다. 다음으로 선별된 SNP들을 복합적으로 고려하여, 계수추정 및 변수선택을 동시에 수행하는 벌점화를 통한 축소추정법을 실시한다. 최종적으로

복합 SNP 분석단계에서 계산한 점수 (measure)를 사용하여 선택된 SNP들에 순서를 부여한다. 기존 연구에서는 SNP의 점수 계산 방법으로, 축소추정법의 회귀계수추정 값과 selection stability를 주로 사용하였다. 그러나 축소추정법을 이용한 회귀계수 추정은 편의 (bias)가 있고, selection stability는 한정된 수의 자료에서는 신뢰할만한 결과를 도출하지 못한다는 단점이 있다.

본 논문은 고차원의 유전체 SNP 자료의 복합적인 영향을 고려하여 복합질환의 예측력을 높일 수 있는 SNP 지표의 발굴 방법에 대하여 고찰하였다. 다단계 접근 방법을 이용한 기존 연구에서 제시하였던 다양한 SNP의 점수 계산 방법의 장 · 단점을 논의하고, 붓스트랩 (bootstrap)과 순열치환 (permutation)과 같은 리샘플링 (resampling) 방법을 바탕으로 새로운 점수 계산 방법을 제안하였다. 다양한 점수 계산 방법들을 사용하여 발굴한 SNP 지표의 복합질환 예측력을 비교하기 위하여, Wellcome Trust Case Control Consortium (WTCCC) 양극성 장애 (bipolar disease: BD)자료에서 얻어진 4,806명의 표본과 354,022개의 SNP을 사용하였다. Elastic-net 벌점함수를 사용한 다단계 분석을 수행하였고, support vector machine (SVM), random forest (RF), elastic-net (EN), 선형 판별 분석 (linear discriminant analysis :LDA)와 같은 다양한 예측 방법을 사용하였다. 다양한 SNP 점수 계산 방법의 평가 및 비교를 위해 area under the curve (AUC)를 사용하였다. 본 논문에서 제안한 SNP 점수 계산 방법은 단일 SNP 분석의 유의 확률을 이용한 SNP 점수 계산 방법보다 높은 예측력을 보였다. 또한, Elastic-net 벌점함수만을 사용한 변수선택법과 비교하여, 조율모수 λ가 작은 값을 가질 경우에 LDA,

EN, SVM 예측 모형에서 보다 높은 예측력을 보이는 SNP을 선택하는 경향이 있음을 확인하였다.

주요어 : 전장유전체 연관분석; Welcome trust case control consotrium (WTCCC); 양극성 장애; joint selection via elastic net; permuted p-value

학 번 : 2011 - 20245