G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(4):149-159
https://doi.org/10.5808/GI.2016.14.4.149

ORIGINAL ARTICLE

Prediction of Quantitative Traits Using Common Genetic Variants: Application to Body Mass Index

Sunghwan Bae[1,2], Sungkyoung Choi[1,2], Sung Min Kim[2], Taesung Park[1,2,3]*

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea,
[2]Bioinformatics and Biostatistics Lab, Seoul National University, Seoul 08826, Korea,
[3]Department of Statistics, Seoul National University, Seoul 08826, Korea

With the success of the genome-wide association studies (GWASs), many candidate loci for complex human diseases have been reported in the GWAS catalog. Recently, many disease prediction models based on penalized regression or statistical learning methods were proposed using candidate causal variants from significant single-nucleotide polymorphisms of GWASs. However, there have been only a few systematic studies comparing existing methods. In this study, we first constructed risk prediction models, such as stepwise linear regression (SLR), least absolute shrinkage and selection operator (LASSO), and Elastic-Net (EN), using a GWAS chip and GWAS catalog. We then compared the prediction accuracy by calculating the mean square error (MSE) value on data from the Korea Association Resource (KARE) with body mass index. Our results show that SLR provides a smaller MSE value than the other methods, while the numbers of selected variables in each model were similar.

Keywords: body mass index, clinical prediction rule, genome-wide association study, penalized regression models, variable selection

Introduction

With the development of genotyping technologies, many disease-related genetic variants have been verified by genome-wide association studies (GWASs). Diagnosis and disease risk prediction from the utilization of the genetic variants have improved even further [1]. Direct-to-consumer genetic companies, such as 23andME (http://www.23andme.com/) and Pathway Genomics (https://www.pathway.com/), provide personal genome information services. For example, the *BRCA1* and *BRCA2* genes play important roles in breast cancer diagnosis and clinical treatment [2, 3]. While several disease prediction studies have been conducted using disease-related genetic variants, there are some limitations to disease risk prediction. It becomes difficult to construct a disease risk prediction model, because there are typically a larger number of genetic variants than the number of individuals in the "large p small n" problem. Also, the effect size of genetic variants for most complex human diseases is small, and missing heritability exists [4]. Moreover, some loss of statistical power to identify significant associations is caused by the correlating single-nucleotide polymorphisms (SNPs) due to linkage disequilibrium (LD) [5]. Multicollinearity due to high LD among SNPs causes high variance of coefficient estimates. In order to solve these issues, various statistical approaches have been recently proposed.

Initially, a gene score (GS) was computed using statistical models for disease risk prediction [6-8]. These risk prediction models were created from GSs by summing up the marginal effect of each disease-associated genetic variant. Several studies have shown that GS is useful for risk prediction [9]. However, the accuracy of the risk prediction is poor when joint effects exist between multiple genetic variants [10, 11].

Building a risk prediction model using multiple SNPs is an effective way to improve disease risk prediction. Multiple logistic regression (MLR) is one of the typical traditional approaches. Several studies have shown the usefulness of an MLR-based approach for creating disease risk prediction

Received November 21, 2016; Revised December 6, 2016; Accepted December 6, 2016

***Corresponding author:** Tel: +82-2-880-8924, Fax: +82-2-883-6144, E-mail: tspark@stats.snu.ac.kr

models [12-14]. However, the parameter estimation of MLR becomes unstable, and the predictive power of the risk prediction model decreases if there is high LD among SNPs.

In order to solve the "large p and small n" problem, many penalized regression approaches, like ridge [15-17], least absolute shrinkage and selection operator (LASSO) [18], and Elastic-Net (EN) [19], have been proposed. For high-dimensional data, these penalized approaches have several advantages in variable selection, as well as in prediction, over non-penalized approaches. For example, several researchers showed that the utilization of a large amount of SNPs with penalized regression approaches improves the accuracy of Crohn's disease and bipolar disorder risk prediction [20, 21].

It is important to build a risk prediction model that pertains to discrete variables, such as disease diagnosis. It is also important to make predictions based on continuous variables, such as human health-related outcomes. When using medicines to treat diseases, we can use genetic information to calculate the dosage, in addition to basic physical information, such as height and weight. For example, there is a prediction model for warfarin responsiveness that was made with multivariate linear regression [22]. We can apply such a model directly to disease treatment.

In this study, we focus on the prediction of quantitative traits using common genetic variants. We systematically compared the performance of prediction models through real data from the Korea Association Resource (KARE). We first selected the prediction variables using statistical methods, such as stepwise linear regression (SLR), LASSO, and EN. We then constructed commonly used risk prediction models, such as SLR, LASSO, and EN. Finally, we compared the predictive accuracy by calculating the mean square error (MSE) value for predicting body mass index (BMI). Overall, our results show that LASSO and SLR provide the smallest MSE value among the compared methods.

Table 1. Demographic variables for KARE cohort

Variable	Total
No. of samples	8,838
Sex (male [%]/female [%])	4,179 (47.3)/4,659 (52.7)
Area (Anseong/Ansan)	4,201/4,637
Age (mean ± SD, yr)	52.22 ± 8.92
BMI (mean ± SD, kg/m^2)	24.60 ± 3.12

KARE, Korea Association Resource; BMI, body mass index.

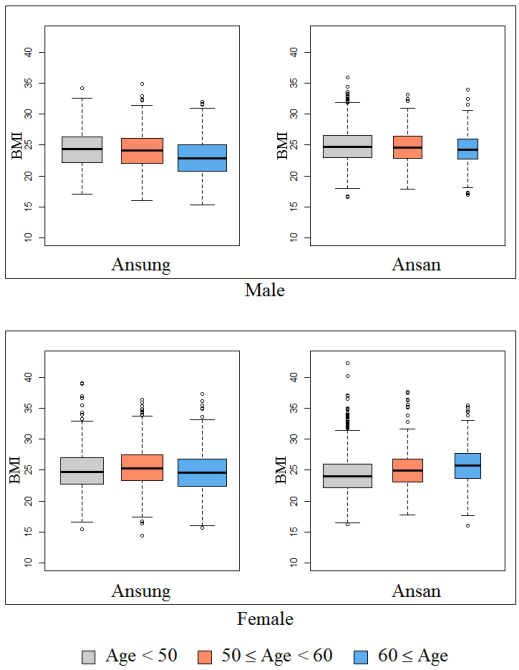

Fig. 1. Box plots of body mass index (BMI) for the given demographic variables.

Table 2. List of the SNP sets

SNP-set	Description	No. of SNPs (GWAS catalog)	No. of SNPs (KARE)	No. of total SNPs
ASIAN-100	GWAS catalog + KARE	16	84	100
KOREAN-100	GWAS catalog + KARE	1	99	100
ALL-200	GWAS catalog + KARE	136	64	200
ASIAN-200	GWAS catalog + KARE	16	184	200
KOREAN-200	GWAS catalog + KARE	1	199	200
GWAS-ALL	Only reported SNPs in GWAS catalog	136	-	136
GWAS-ASIAN	Only reported SNPs in GWAS catalog	16	-	16

SNP, single nucleotide polymorphism; GWAS, genome-wide association study; KARE, Korea Association Resource; ASIAN-100, GWAS catalog (Asia) + single-SNP analysis; KOREAN-100, GWAS catalog (Korea) + single-SNP analysis; ALL-200, GWAS catalog (All) + single-SNP analysis; ASIAN-200, GWAS catalog (Asia) + single-SNP analysis; KOREAN-200, GWAS catalog (Korea) + single-SNP analysis; GWAS-ALL, GWAS catalog (All); GWAS-ASIAN, GWAS catalog (Asia).

Methods

Data

The KARE project, which began in 2007, is an Anseong and Ansan regional society-based cohort. After applying SNP quality control criteria—Hardy-Weinberg equilibrium $p < 10^{-06}$, genotype call rates < 95%, and minor allele frequency < 0.01—352,228 SNPs were utilized for analysis. Also, after eliminating 401 samples with call rates less than 96%, 11 contaminated samples, 41 gender-inconsistent samples, 101 serious concomitant illness samples, 608 cryptic-related samples, and 4 samples with missing phenotype, 8,838 participants were analyzed [23]. Table 1 summarizes the demographic information. In addition, Fig. 1 shows box plots of BMI for the given demographic variables.

Statistical analysis

We selected SNPs from the KARE data analysis based on single-SNP analysis and collected SNPs in the GWAS catalog [24]. Then, we performed two steps to make quantitative prediction models. First, we selected the variables by using SLR, LASSO, and EN and then built quantitative prediction models by using the same methods.

SNP sets

First, based on three different populations—overall population, Asian-only population, and Korean-only population—we collected the SNPs registered in the GWAS catalog for BMI. Second, the SNPs were selected by single-SNP analysis using linear regression with adjustments for sex, age, and area. We chose the SNPs based on the p-values. We considered the following seven SNP sets:

(1) **ASIAN-100** (GWAS catalog [Asia] + Single-SNP analysis, number of SNPs = 100)
(2) **KOREAN-100** (GWAS catalog [Korea] + single-SNP analysis, number of SNPs = 100)
(3) **ALL-200** (GWAS catalog [All] + single-SNP analysis, number of SNPs = 200)
(4) **ASIAN-200** (GWAS catalog [Asia] + single-SNP analysis, number of SNPs = 200)
(5) **KOREAN-200** (GWAS catalog [Korea] + single-SNP analysis, number of SNPs = 200)
(6) **GWAS-ALL** (GWAS catalog [All], number of SNPs = 136)
(7) **GWAS-ASIAN** (GWAS catalog [Asia], number of SNPs = 16)

Step 1: Variable selection

In the KARE data, out of 8,838 individuals, we randomly selected 1,767 for test sets and composed the training set with the rest of the 7,071 participants. We selected SNPs

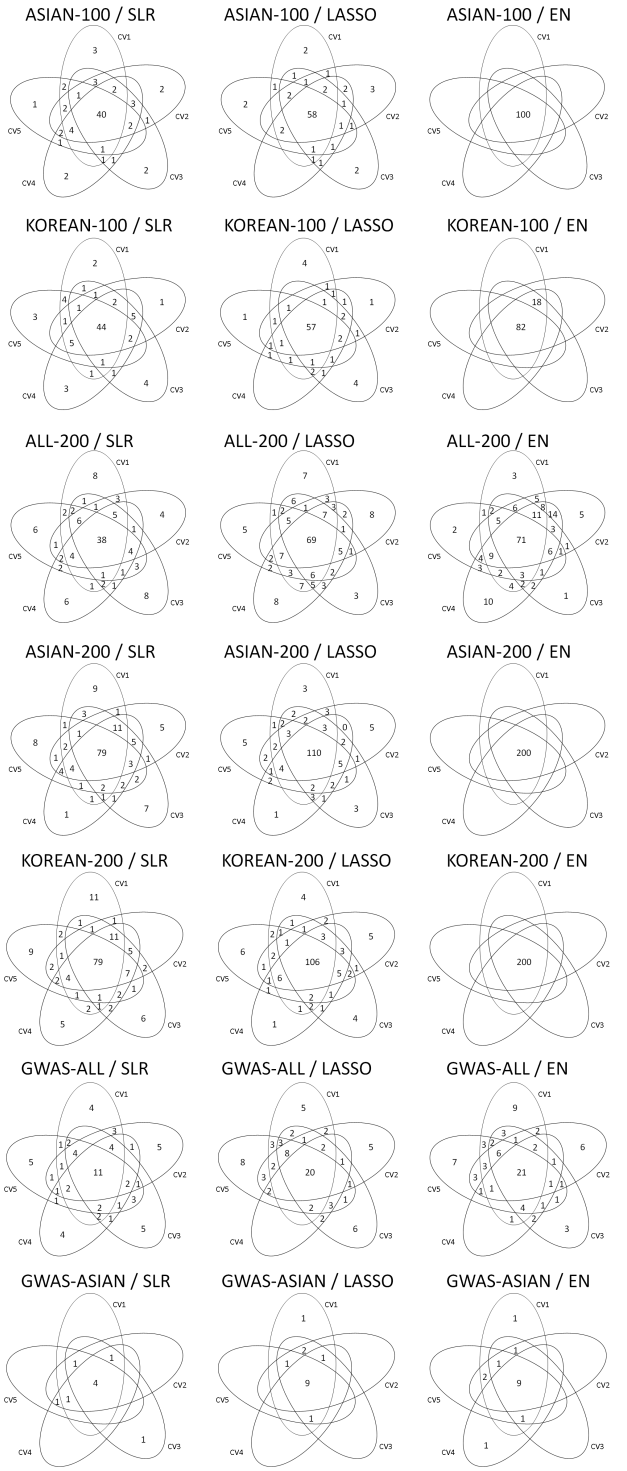

Fig. 2. Venn diagrams give us shared parts from 5-fold CV by variables selection methods. CV, cross-validation; ASIAN-100, genome-wide association study (GWAS) catalog (Asia) + single-single-nucleotide polymorphism (SNP) analysis; KOREAN-100, GWAS catalog (Korea) + single-SNP analysis; ALL-200, GWAS catalog (All) + single-SNP analysis; ASIAN-200, GWAS catalog (Asia) + single-SNP analysis; KOREAN-200, GWAS catalog (Korea) + single-SNP analysis; GWAS-ALL, GWAS catalog (All); GWAS-ASIAN, GWAS catalog (Asia); SLR, stepwise linear regression; LASSO, least absolute shrinkage and selection operator; EN, Elastic-Net.

using 5-fold cross-validation (CV) of the training set. In this case, we used SLR, LASSO, and EN to select SNPs.

The SLR model is one of the most widely used models. Let y_i be a quantitative phenotype for subject $i = 1, \cdots, n$; x_{ij} be the value of SNP $j = 1, \cdots, p$ for subject i; code be 0, 1, and 2 for the number of minor alleles; and ε_i be the error term for subject i. The SLR model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \gamma_1 \text{sex}_i + \gamma_2 \text{age}_i + \gamma_3 \text{area}_i + \varepsilon_i,$$

where β_0 and β_j are the intercept and effect sizes of SNPs, respectively. γ_1, γ_2, and γ_3 represent the sex, age, and area of the i-th individual, respectively. Variable selection was performed by a MSE-based stepwise procedure. The stepwise procedure was performed using the R package "MASS" [25].

The LASSO and EN estimates of β were obtained by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi} - \gamma_1 \text{sex}_i - \gamma_2 \text{age}_i - \gamma_3 \text{area}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

and

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi} - \gamma_1 \text{sex}_i - \gamma_2 \text{age}_i - \gamma_3 \text{area}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2,$$

respectively. The tuning parameters λ_1 and λ_2 are estimated using CV. The penalized methods were performed using the R package "glmnet" [26].

Then, we defined five groups.

(1) **Group 1** (consists of SNPs that appeared at least one time in the 5-fold CV)
(2) **Group 2** (consists of the SNPs that appeared at least two times in the 5-fold CV)
(3) **Group 3** (consists of the SNPs that appeared at least three times in the 5-fold CV)
(4) **Group 4** (consists of the SNPs that appeared at least four times in the 5-fold CV)
(5) **Group 5** (consists of the SNPs that appeared in all 5-fold CVs)

Table 3. The number of overlapping SNPs selected by 5-fold CV for each variable selection method

SNP-sets	Variable selection method	Group 1	Group 2	Group 3	Group 4	Group 5
ASIAN-100	SLR	76	66	61	50	40
	LASSO	86	77	71	66	58
	EN	100	100	100	100	100
KOREAN-100	SLR	82	69	62	55	44
	LASSO	87	77	72	63	57
	EN	100	100	100	100	82
ALL-200	SLR	113	81	67	58	38
	LASSO	174	143	119	99	69
	EN	185	164	134	105	71
ASIAN-200	SLR	156	126	115	100	79
	LASSO	171	154	141	127	110
	EN	200	200	200	200	200
KOREAN-200	SLR	162	128	115	102	79
	LASSO	166	146	136	123	106
	EN	200	200	200	200	200
GWAS-ALL	SLR	67	44	33	25	11
	LASSO	82	58	45	32	20
	EN	85	60	45	35	2
GWAS-ASIAN	SLR	9	8	8	7	4
	LASSO	16	14	14	12	9
	EN	16	14	14	11	9

SNP, single nucleotide polymorphism; CV, cross-validation; ASIAN-100, genome-wide association study (GWAS) catalog (Asia) + single-SNP analysis; SLR, stepwise linear regression; LASSO, least absolute shrinkage and selection operator; EN, Elastic-Net; KOREAN-100, GWAS catalog (Korea) + single-SNP analysis; ALL-200, GWAS catalog (All) + single-SNP analysis; ASIAN-200, GWAS catalog (Asia) + single-SNP analysis; KOREAN-200, GWAS catalog (Korea) + single-SNP analysis; GWAS-ALL, GWAS catalog (All); GWAS-ASIAN, GWAS catalog (Asia).

Table 4. MSE values from test dataset

SNP-set	Variable selection methods	Group	Prediction method			
			Only used covariates	SLR	LASSO	EN
ASIAN-100	LASSO	1	10.24	12.81	12.06	11.15
		2	10.24	13.08	12.35	13.02
		3	10.24	13.08	10.77	10.97
		4	10.24	12.81	11.72	11.40
		5	10.24	**9.64**	9.66	9.69
	EN	1	10.24	**9.64**	9.70	9.87
		2	10.24	**9.64**	9.70	9.87
		3	10.24	**9.64**	9.70	9.87
		4	10.24	**9.64**	9.70	9.87
		5	10.24	**9.64**	9.70	9.87
	SLR	1	10.24	19.99	18.05	12.52
		2	10.24	24.72	16.24	19.23
		3	10.24	16.94	16.39	14.67
		4	10.24	15.30	14.44	11.61
		5	10.24	9.75	9.76	9.77
KOREAN-100	LASSO	1	10.24	12.50	13.02	12.40
		2	10.24	12.04	13.99	11.51
		3	10.24	12.47	12.66	11.72
		4	10.24	10.37	14.60	13.55
		5	10.24	9.69	9.70	9.72
	EN	1	10.24	17.78	9.73	13.41
		2	10.24	17.78	9.73	13.41
		3	10.24	17.78	9.73	13.41
		4	10.24	17.78	9.73	13.41
		5	10.24	9.66	9.71	9.77
	SLR	1	10.24	20.47	13.75	12.42
		2	10.24	20.47	13.25	12.28
		3	10.24	18.25	17.19	15.87
		4	10.24	17.60	14.99	11.11
		5	10.24	9.76	9.76	9.77
ALL-200	LASSO	1	10.24	14.84	11.75	12.59
		2	10.24	15.55	12.79	13.20
		3	10.24	15.60	15.48	12.98
		4	10.24	12.86	13.85	12.24
		5	10.24	9.86	9.91	9.92
	EN	1	10.24	15.02	11.59	12.06
		2	10.24	16.10	12.73	12.64
		3	10.24	11.81	13.89	12.86
		4	10.24	13.75	12.57	11.80
		5	10.24	9.87	9.91	9.93
	SLR	1	10.24	16.03	20.81	13.01
		2	10.24	16.14	17.97	18.12
		3	10.24	20.11	18.24	18.42
		4	10.24	20.00	17.80	18.01
		5	10.24	9.84	9.85	9.86
ASIAN-200	LASSO	1	10.24	23.87	16.36	12.95
		2	10.24	13.13	19.14	13.72
		3	10.24	16.07	17.55	17.90
		4	10.24	15.46	14.29	12.47
		5	10.24	9.67	9.73	9.74
	EN	1	10.24	9.80	9.87	10.21

Table 4. Continued 1

SNP-set	Variable selection methods	Group	Prediction method			
			Only used covariates	SLR	LASSO	EN
		2	10.24	9.80	9.87	10.21
		3	10.24	9.80	9.87	10.21
		4	10.24	9.80	9.87	10.21
		5	10.24	9.80	9.87	10.21
	SLR	1	10.24	30.37	15.40	16.63
		2	10.24	24.02	21.30	12.51
		3	10.24	23.32	26.35	25.17
		4	10.24	21.84	17.56	18.25
		5	10.24	9.87	9.86	9.87
KOREAN-200	LASSO	1	10.24	23.22	23.37	16.22
		2	10.24	13.33	17.80	13.93
		3	10.24	16.86	15.26	15.54
		4	10.24	18.91	14.15	12.88
		5	10.24	9.71	9.78	9.78
	EN	1	10.24	9.82	9.86	10.23
		2	10.24	9.82	9.86	10.23
		3	10.24	9.82	9.86	10.23
		4	10.24	9.82	9.86	10.23
		5	10.24	9.82	9.86	10.23
	SLR	1	10.24	38.31	18.12	13.61
		2	10.24	37.07	18.31	15.04
		3	10.24	29.48	18.61	16.80
		4	10.24	18.60	16.16	15.84
		5	10.24	9.93	9.93	9.92
GWAS-ALL	LASSO	1	10.24	10.79	10.90	10.84
		2	10.24	10.99	11.23	10.84
		3	10.24	10.99	10.88	10.77
		4	10.24	10.52	10.62	10.56
		5	10.24	10.27	10.27	10.27
	EN	1	10.24	10.84	10.67	10.83
		2	10.24	10.92	11.00	10.66
		3	10.24	10.92	11.24	11.02
		4	10.24	11.00	10.94	10.86
		5	10.24	10.26	10.26	10.26
	SLR	1	10.24	12.21	10.67	10.62
		2	10.24	11.95	11.69	10.74
		3	10.24	11.50	10.92	10.41
		4	10.24	11.36	11.20	10.63
		5	10.24	10.26	10.26	10.25
GWAS-ASIAN	LASSO	1	10.24	10.12	10.23	10.23
		2	10.24	10.12	10.23	10.43
		3	10.24	10.12	10.23	10.43
		4	10.24	10.12	10.42	10.43
		5	10.24	10.12	10.13	10.13
	EN	1	10.24	10.12	10.22	10.35
		2	10.24	10.12	10.35	10.36
		3	10.24	10.12	10.35	10.36
		4	10.24	10.12	10.32	10.33
		5	10.24	10.12	10.13	10.13
	SLR	1	10.24	10.35	10.37	10.36
		2	10.24	10.35	10.35	10.34

Table 4. Continued 2

SNP-set	Variable selection methods	Group	Prediction method			
			Only used covariates	SLR	LASSO	EN
		3	10.24	10.35	10.35	10.34
		4	10.24	10.18	10.18	10.18
		5	10.24	10.17	10.17	10.17

MSE, mean square error; SNP, single nucleotide polymorphism; SLR, stepwise linear regression; LASSO, least absolute shrinkage and selection operator; EN, Elastic-Net; ASIAN-100, GWAS catalog (Asia) + single-SNP analysis; KOREAN-100, GWAS catalog (Korea) + single-SNP analysis; ALL-200, GWAS catalog (All) + single-SNP analysis; ASIAN-200, GWAS catalog (Asia) + single-SNP analysis; KOREAN-200, GWAS catalog (Korea) + single-SNP analysis; GWAS-ALL, GWAS catalog (All); GWAS-ASIAN, GWAS catalog (Asia).

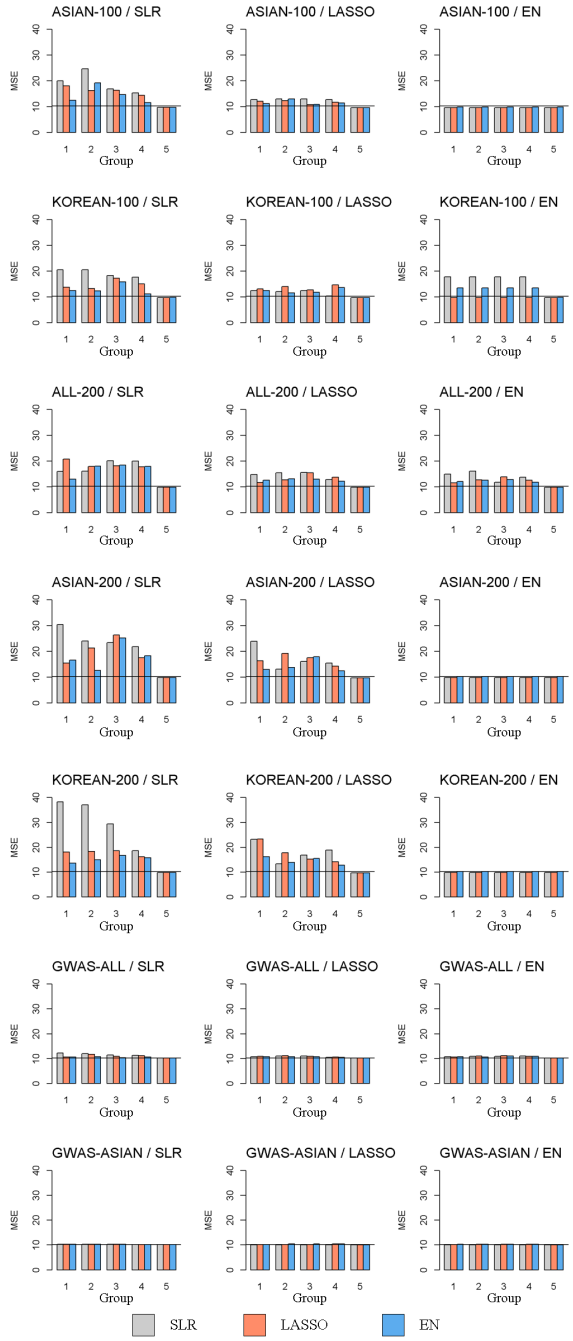

Fig. 3. Each set by MSE value, x-axis are the number of CV containing the selected variable. Group 1, 5 is a model from variables of the union of CV and of the intersection of CV, respectively. The gray bar indicates the SLR, the orange bar indicates the LASSO, the blue bar indicates the EN and the black line is MSE value of 10.24 from the prediction model using only covariates. MSE, mean square error; CV, cross-validation; ASIAN-100, genome-wide association study (GWAS) catalog (Asia) + single-single-nucleotide polymorphism (SNP) analysis; KOREAN-100, GWAS catalog (Korea) + single-SNP analysis; ALL-200, GWAS catalog (All) + single-SNP analysis; ASIAN-200, GWAS catalog (Asia) + single-SNP analysis; KOREAN-200, GWAS catalog (Korea) + single-SNP analysis; GWAS-ALL, GWAS catalog (All); GWAS-ASIAN, GWAS catalog (Asia); SLR, stepwise linear regression; LASSO, least absolute shrinkage and selection operator; EN, Elastic-Net.

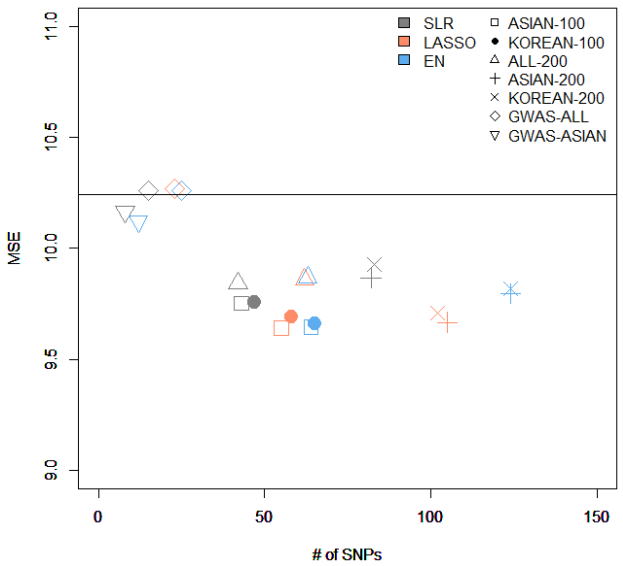

Fig. 4. The comparison of the results from variables selected by different methods and from creating a model using stepwise. MSE, mean square error; SNP, single-nucleotide polymorphism; ASIAN-100, genome-wide association study (GWAS) catalog (Asia) + single-SNP analysis; KOREAN-100, GWAS catalog (Korea) + single-SNP analysis; ALL-200, GWAS catalog (All) + single-SNP analysis; ASIAN-200, GWAS catalog (Asia) + single-SNP analysis; KOREAN-200, GWAS catalog (Korea) + single-SNP analysis; GWAS-ALL, GWAS catalog (All); GWAS-ASIAN, GWAS catalog (Asia); SLR, stepwise linear regression; LASSO, least absolute shrinkage and selection operator; EN, Elastic-Net.

Step 2: Quantitative prediction

To build a quantitative prediction model, we used the same prediction methods that were applied for the variable selection step for the comparison of these three methods in the variable selection and quantitative prediction. Each prediction model was created by using 7,071 training individuals via 5-fold CV. To compare the performance of the quantitative prediction models, we calculated the MSE by applying each quantitative prediction model using the test set (n = 1,767).

Results

To create the SNP sets associated with BMI, single-SNP analysis was performed by linear regression with adjustments for sex, age, and area. As shown in Supplementary Fig. 1, we found one significant SNP (rs17178527) after Bonferroni correction (1.45×10^{-07}). rs17178527 of *LOC729076* has been reported as BMI-associated SNP in previous GWASs [23, 27]. In addition, Supplementary Table 1 shows the results of the single-SNP analysis with p-values less than 5.00×10^{-05}. The SNPs that were reported to be associated with BMI in the GWAS catalog are summarized in Supplementary Table 2. Seven SNP sets are summarized in Table 2.

Step 1: Variable selection

Variable selection in each SNP set was performed via 5-fold CV of the training set. Fig. 2 shows the overlapping number of selected SNPs by the variable selection methods. In addition, Table 3 provides more detailed information. Overall, SLR selected fewer SNPs than LASSO and EN. All SNPs were selected when EN was used in ASIAN-100, ASIAN-200, and KOREAN-200.

Step 2: Quantitative prediction

We made quantitative prediction models based on SLR, LASSO, and EN using the entire training dataset. Then, the MSE was calculated by applying the quantitative prediction models to the test dataset. Table 4 and Fig. 3 show the performance of each quantitative prediction model in the test dataset. The model using only covariates yielded an MSE value of 10.24. As can be seen from Fig. 3, the prediction model created from Group 5 yielded the smallest MSE. Fig. 4 describes the comparison results between the numbers of SNPs and MSEs from the prediction models using SLR.

Among all sets, the case that used LASSO to select variables and SLR to create the model showed the smallest MSE value of 9.64 in ASIAN-100, with 51 SNPs. Among the 51 SNPs of LASSO-SLR with one set from ASIAN-100, 28 SNPs were mapped to genes (Table 5). Some genes, such as *FTO*, *GP2*, *AKAP6*, *ANKS1B*, *ADCY3*, and *ADCY8*, have been reported to be associated with BMI [28-33].

Discussion

In this study, we used statistical methods (SLR, LASSO, and EN) to select variables and build quantitative prediction models. Then, we compared the performance of the quantitative prediction models by each SNP set (ASIAN-100, KOREAN-100, ALL-200, ASIAN-200, KOREAN-200, GWAS-ALL, and GWAS-ASIAN). As a result, the performance of the prediction models using the GWAS catalog and KARE data was better than that of the prediction models using only SNPs reported in the GWAS catalog. For the case that selected variants using LASSO in ASIAN-100 and created a prediction model using SLR, the MSE value was the smallest, 9.64. At this time, the number of SNPs was 51. Also, for the model with the fewest SNPs, we selected variables using SLR from ALL-200 and created a model using SLR. The number of SNPs was 38, and the MSE value was 9.84. Through the 5-fold CV, we developed a quantitative prediction model. After calculating MSE from groups 1 to 5, when assembled with SNPs that were included in all CVs, the resulting values of MSE were small. However, when a different group was

Table 5. Development of LASSO-SLR prediction model with one set from ASIAN-100 for predicting BMI

SNP	β	Region	Gene	SNP	β	Region	Gene
rs17411146	−0.41	Upstream	-	rs11984203	0.19	Intron	*NUP205*
rs4121165	−0.16	Intron	*FAM73A*	rs2726602	−0.22	Downstream	*TOX*
rs12142366	0.27	Intron	*ELTD1*	rs2721109	−0.18	Upstream	-
rs17130257	−0.26	Downstream	-	rs16904384	0.71	Intron	*ADCY8*
rs4081366	0.16	Downstream	-	rs10961819	0.18	Upstream	-
rs527248	0.21	downstream	-	rs4287251	0.64	Intron	-
rs1281296	−0.32	Downstream	*ZNF648*	rs11000212	0.28	Intron	*ASCC1*
rs12092943	0.19	Intron	*PIK3C2B*	rs11193517	−0.26	Downstream	-
rs6545814	0.12	Intron	*ADCY3*	rs11030104	−0.11	Intron	-
rs12615642	0.11	Intron	-	rs652722	−0.13	Intron	-
rs10207849	0.18	Upstream	-	rs7108746	−0.20	intron	-
rs11893160	−0.29	Intron	*FHL2*	rs7107562	0.25	downstream	-
rs7424822	0.28	Intron	*THSD7B*	rs402590	0.50	Intron	*ANO2*
rs9839685	0.52	Intron	*ATP2B2*	rs4272863	−0.30	Intron	*AMN1*
rs1399903	0.20	Downstream	-	rs17092358	0.16	Downstream	-
rs4626221	−0.22	Intron	-	rs2373011	0.09	Intron	*ANKS1B*
rs1491332	−0.21	Downstream	-	rs12229654	−0.15	Upstream	*CUX2*
rs10056782	0.18	Intron	*PPP2R2B*	rs2296189	−0.21	CDS	*FLT1*
rs6893893	−0.20	Intron	*ATP10B*	rs7995818	−0.11	Downstream	-
rs792965	−0.43	Intron	*ERGIC1*	rs9569190	0.33	Downstream	-
rs3857596	0.21	Downstream	-	rs10483416	0.22	Intron	*AKAP6*
rs1342644	0.16	Intron	*PEX7*	rs12597579	−0.11	Downstream	*GP2*
rs17178527	−0.28	-	-	rs9939609	0.31	Intron	*FTO*
rs4509217	0.61	Intron	*HECW1*	rs633265	0.22	Upstream	-
rs9987062	0.28	Downstream	*C7orf66*	rs4802919	0.16	Upstream	*ZNF480*
rs2188187	−0.29	Intron	*GRM8*				

LASSO, least absolute shrinkage and selection operator; SLR, stepwise linear regression; ASIAN-100, GWAS catalog (Asia) + single-SNP analysis; BMI, body mass index; SNP, single-nucleotide polymorsphism.

used, the MSE value was bigger than when using the covariates to build the model. Therefore, with CV, when using SNPs that match each of their CVs, the efficiency of their quantitative prediction model was high. In the variable selection, SLR performed better than other methods. SLR selected fewer SNPs than the other methods in all SNP sets while providing smaller MSEs. It seems that LASSO and EN tended to select SNPs with little contribution to BMI. For further research, we plan to perform simulation studies and a real-data analysis with other continuous traits.

There are many ways to extend the analysis of quantitative prediction studies. First, along with the application of recently developed methods, such as bootstrapping methods [34, 35], we will continue to explore new ways to develop more prediction models. Second, the incorporation of rare variants can improve the performance of a quantitative prediction model. Advanced sequencing technology has made it possible to investigate the role of common and rare variants in complex disease risk prediction. Additionally, we can use biological information while choosing the variables. By using single-SNP analysis, we can use gene or pathway information to find useful SNPs [36], and from here, we can assemble an SNP set by adding an SNP list from the pathways related to the disease of interest.

Supplementary materials

Supplementary data including two tables and one figure can be found with this article online http://www.genominfo.org/src/sm/gni-14-149-s001.pdf.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI15C2165), and the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation. The GWAS chip data were supported by bioresources from the National Biobank of Korea, the Centers for Disease Control and

Prevention, Republic of Korea (4845-301, 4851-302 and -307).

References

1. Kooperberg C, LeBlanc M, Obenchain V. Risk prediction using genome-wide association studies. *Genet Epidemiol* 2010;34: 643-652.
2. Futreal PA, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S, *et al*. *BRCA1* mutations in primary breast and ovarian carcinomas. *Science* 1994;266:120-122.
3. Lancaster JM, Wooster R, Mangion J, Phelan CM, Cochran C, Gumbs C, *et al*. *BRCA2* mutations in primary breast and ovarian cancers. *Nat Genet* 1996;13:238-240.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-753.
5. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;6:109-118.
6. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, *et al*. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009;460:748-752.
7. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet Epidemiol* 2011;35:506-514.
8. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 2009;18:3525-3531.
9. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet* 2008;17:R166-R173.
10. Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, *et al*. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med* 2006;3:e374.
11. van der Net JB, Janssens AC, Sijbrands EJ, Steyerberg EW. Value of genetic profiling for the prediction of coronary heart disease. *Am Heart J* 2009;158:105-110.
12. Lindström S, Schumacher FR, Cox D, Travis RC, Albanes D, Allen NE, *et al*. Common genetic variants in prostate cancer risk prediction: results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). *Cancer Epidemiol Biomarkers Prev* 2012;21:437-444.
13. Jostins L, Barrett JC. Genetic risk prediction in complex disease. *Hum Mol Genet* 2011;20:R182-R188.
14. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, *et al*. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* 2010;362:986-993.
15. Hoerl AE. Ridge regression. *Biometrics* 1970;26:603.
16. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. *Technometrics* 1970;12:69-82.
17. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970;12:55-67.
18. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* 1996;58:267-288.
19. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* 2005;67:301-320.
20. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, *et al*. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet* 2013;92:1008-1012.
21. Austin E, Pan W, Shen X. Penalized regression and risk prediction in genome-wide association studies. *Stat Anal Data Min* 2013;6:315-328.
22. Cha PC, Mushiroda T, Takahashi A, Kubo M, Minami S, Kamatani N, *et al*. Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. *Hum Mol Genet* 2010;19:4735-4744.
23. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
24. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;42:D1001-D1006.
25. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, *et al*. Package 'MASS'. CRAN Repository, 2013. Accessed 2016 Dec 1. Available from: http://cran r-project org/web/packages/MASS/MASS pdf.
26. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1-22.
27. Kim J, Namkung J, Lee S, Park T. Application of structural equation models to genome-wide association analysis. *Genomics Inform* 2010;8:150-158.
28. Wang KS, Liu X, Owusu D, Pan Y, Xie C. Polymorphisms in the *ANKS1B* gene are associated with cancer, obesity and type 2 diabetes. *AIMS Genet* 2015;2:192-203.
29. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, *et al*. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;316:889-894.
30. Wen W, Cho YS, Zheng W, Dorajoo R, Kato N, Qi L, *et al*. Meta-analysis identifies common variants associated with body mass index in east Asians. *Nat Genet* 2012;44:307-311.
31. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, *et al*. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 2012;44: 659-669.
32. Sung YJ, Pérusse L, Sarzynski MA, Fornage M, Sidney S, Sternfeld B, *et al*. Genome-wide association studies suggest sex-specific loci associated with abdominal and visceral fat. *Int J Obes (Lond)* 2016;40:662-674.
33. Stergiakouli E, Gaillard R, Tavaré JM, Balthasar N, Loos RJ, Taal HR, *et al*. Genome-wide association study of height-adjusted BMI in childhood identifies functional variant in *ADCY3*. *Obesity (Silver Spring)* 2014;22:2252-2259.
34. Hall P, Lee ER, Park BU. Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Stat Sin* 2009;19:449-

471.
35. Chatterjee A, Lahiri SN. Bootstrapping Lasso estimators. *J Am Stat Assoc* 2011;106:608-625.
36. Eleftherohorinou H, Wright V, Hoggart C, Hartikainen AL, Jarvelin MR, Balding D, *et al*. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS One* 2009;4:e8068.