저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 석 사 학 위 논 문

**Risk prediction using common and rare
genetic variants
: application to Type 2 diabetes**

빈도가 높은 유전적 변이와 빈도가 낮은 유전적
변이를 이용한 위험 예측
: 제 2 형 당뇨 적용

2018 년 2 월

서울대학교 대학원
협동과정 생물정보학과
배 성 환

**Risk prediction using common and rare
genetic variants**

: application to Type 2 diabetes

by

Sunghwan Bae

A thesis

**Submitted in fulfillment of the requirement
for the degree of Master**

in

Bioinformatics

Interdisciplinary Program in Bioinformatics

College of Natural Sciences

Seoul National University

Feb, 2018

Risk prediction using common and rare genetic variants

: application to Type 2 diabetes

지도교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함

2018 년 2 월

서울대학교 대학원

협동과정 생물정보학과

배 성 환

배성환의 이학석사 학위论문을 인준함

2018 년 2 월

위 원 장 원 성 호 (인)

부위원장 박 태 성 (인)

위 원 이 승 연 (인)

Abstract

Risk prediction using common and rare genetic variants : application to Type 2 diabetes

Sunghwan Bae

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Genome-wide association studies (GWAS) have identified many disease-related common variants. Common genetic variants are being diagnosed and treated. Furthermore, using common genetic variants, there have been several prediction models suggested based on penalized regression or statistical learning methods. However, the common variant is not sufficient to explain the phenotype. One way to

solve this problem is to consider rare variations. This is because rare variants has a large impact on disease. A recent development of next generation sequencing technology (NGS) has identified several disease-related rare genetic variants. However only a few studies have compared predictive models using both common and rare variants. The aim of our study is to compare the performance of prediction models systematically by using common and rare variants from the Whole Exome Sequencing (WES) data of Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Consortium. We first constructed risk prediction models, such as stepwise logistic regression (SLR), least absolute shrinkage and selection operator (LASSO), Elastic-Net (EN) and support vector machine (SVM). We then compared prediction accuracy by calculating the area under the curve (AUC). Our results show that the performance using both common and rare variants was better than using either the common variants only or the rare variants only. Although the AUC values were different depending on the variant sets, the AUC values of SVM prediction models were always larger than those of other prediction models. Among the four rare variant sets, AUC value was larger at pty_ns set.

Keywords: Whole exome sequencing (WES), Risk prediction model, Type 2 diabetes (T2D), Penalized regression methods, Stepwise selection, Support vector machine (SVM)

Student number: 2015-20506

Contents

Abstract	i
Contents	iv
List of Figures	v
List of Tables	vi
1 Introduction	1
2 Material	5
3 Methodology	7
4 Results	15
5 Discussion	24
Bibliography	26
Abstract (Korean)	28

List of Figures

Figure 1 The overall analysis scheme	8
Figure 2 Result of each variant combination and each method using ptv set	19
Figure 3 Result of each variant combination and each method using ptv_ms set	20
Figure 4 Result of each variant combination and each method using ptv_ns_b set	20
Figure 5 Result of each variant combination and each method using ptv_ns_s set	21
Figure 6 Results for KAR and T2D-GENES data. Each bar represents one of fourteen SNP data sets.....	22

List of Tables

Table 1 Demographic variables for T2D-GENES.	5
Table 2 Summary for rare variant sets.	8
Table 3 The number of genes selected by each variant list and collapsed method.	15
Table 4 AUC of the test data set	17

Chapter 1

Introduction

The success of genome-wide association studies (GWAS) has resulted in identification of many disease-related common genetic variants. With the use of genetic variation, diagnostic, treatment and disease risk predictions have been further improved [1]. For example, 23andME (<http://www.23andme.com/>) and Pathway Genomics (<https://www.pathway.com/>) provide personal genome information services. Several disease prediction studies have been conducted using disease-related genetic variants, but there are some limitations to disease risk prediction. First there is the “large p small n” problem. This means that there are typically a larger number of genetic variants than the number of individuals. Second, because of linkage disequilibrium (LD) between genetic variants, statistical power drops in identifying significant association [2]. The high variance of the coefficient estimates is caused by the multicollinearity due to the high LD among the genetic variants.

Various statistical approaches have recently been proposed to address these problems. Penalized methods such as ridge [3-5], least absolute shrinkage and selection operator (LASSO) [6] and Elastic-Net (EN) [7] have been proposed to solve the “large p small n” problem. The approach of penalized method in high dimensional data has more advantages than non-penalized methods in genetic variants selection. For example, several researches showed that the utilization of a large number of single-nucleotide polymorphisms (SNPs) with penalized regression approaches improves the accuracy of Crohn’s disease and inflammatory bowel disease prediction [1, 8]. In addition, data mining approaches have been widely used to improve the risk prediction performance. In particular, support vector machine (SVM) [9, 10] has been shown to outperform the other classification algorithms [11] and to have a good performance in prediction using a large number of SNPs [12].

Most of these prediction models have used the common variants. However, there are many reasons for small effect size of the common variant and one of them is missing heritability [13]. On the other hand, recent studies suggest that rare variants plays an important role in the complex diseases having large effect sizes. For example, rare variants have been associated with autism, atherosclerosis and mental

retardation [14-17]. From an evolutionary standpoint, rare variants are more harmful than common variants because they are under less negative selective pressures [18].

Our earlier work compared the performance of risk prediction models using common variants [19]. In our earlier work, we first built the prediction model by combining of variants selection and prediction methods for the type 2 diabetes (T2D) using the Korean Association Resource (KARE) data. We confirmed that prediction models incorporating both demographic and genetic variables are more accurate than those using demographic variables only. In this current study, we focus on risk prediction using both common and rare variants. We have systematically compared the performance of prediction models using real whole exome sequencing data from Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D GENES) consortium. We compared the performances of three different risk prediction models based on common variants only, rare variants only, and both common and rare variants. We first constructed commonly used risk prediction methods, such as SLR, LASSO, EN and SVM. We then calculated the area under the curve (AUC) values of T2D prediction models to compare their performance.

Our study showed that the SVM prediction model with LASSO selected variables yielded the largest AUC value.

Chapter 2

Material

T2D-GENES consortium is a large scale collaborative study to find genetic variants that affect the risk of T2D. Deep exome sequencing data of 12,940 individuals were collected from five ancestry groups (Europeans, East Asians, South Asians, American Hispanics, and African Americans) [20]. Our study was conducted using 1,086 Korean samples and 1,078 Singaporean samples from T2D-GENES consortium data. A total of 2,892,490 genetic variants are used for analysis after applying the quality control (QC) criteria of Hardy-Weinberg equilibrium (HWE) P-values $< 10^{-6}$ and gene call rate $> 95\%$. Missing genotypes were imputed using Eagle software. Demographic information is summarized in Table 1.

There are two criteria for T2D diagnosis. The first criterion is that fasting plasma glucoses (FPG) larger than or equal to 126 mg/DL, 2-hour postprandial blood glucoses (Glu120) larger than or equal to 200 mg/dL. Another criterion is the treatment of T2D.

Table 1 Demographic variables for T2D-GENES

<i>Information</i>	<i>T2D</i>	<i>Non-T2D</i>
# of sample	526	560
Male/Female	286/240	232/328
Age (year)	53.82±3.58	63.25±7.47
Body Mass Index (BMI)	25.69±3.25	23.72±3.06

There are two criteria for nondiabetic normal. The first criterion is that FPG less than or equal to 100 mg/dL or Glu120 less than or equal to 140 mg/dL. Another criterion is no history of diabetes.

The KARE project that began in 2007 is Ansung and Ansan regional society-based cohort. After applying SNP quality control (QC) criterion of Hardy-Weinberg equilibrium (HWE) $P\text{-value} < 10^{-06}$, genotype call rates $< 95\%$, minor allele frequency (MAF) < 0.01 , total 352,228 SNPs are 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 phenotype missing samples, the total 8,838 participants were analyzed

Chapter 3

Methodology

A. WES data analysis

The common variants were selected from the data base on single-SNP analysis and from GWAS catalog [21]. The rare variants were then selected from the data analysis based on gene based analysis with Combined Multivariate and Collapsing (CMC) method [22]. Next, we constructed risk prediction models from each set of SNP variants combinations using 5-fold cross-validation (CV). The data was randomly split into 5 CV sets, then training set was made with 4 CV sets and test set was made with the other CV set. At step 1, we selected the variants by using the SLR, LASSO and EN. At step 2, we then built the risk prediction models by using the SLR, LASSO, EN and SVM. Fig 1 summarizes the analysis scheme.

Figure 1 The overall analysis scheme

A.1. Variant combinations

First, we collected the common variants registered in GWAS catalog for T2D. Second, the common variants were selected by single-SNP analysis using a logistic regression with adjustment for sex and BMI. We chose the SNPs based on the P-values (1.0×10^{-3}). The rare variants are first collapsed based on the CMC method and weighted CMC method. We then performed gene-based analysis using a logistic regression with adjustment for sex and BMI. We chose the genes based on the P-values (1.0×10^{-2}). For rare variants selection, we considered four rare variant sets based on minor allele frequency (MAF) and functional annotation. We mapped variants to transcripts in Ensembl 66 (GRCh37.66). Table 2 summarizes four rare variant sets.

Table 2 Summary for rare variant sets

ptv: protein truncating variant set (nonsense, frameshift, essential splice site) using annotations from CHAoS v0.6.3, SnpEFF v3.1, and VEP v2.7
ptv_ms: protein-altering variant set (missense, inframeshift, non-essential splice site) in at least one mapped transcript with MAF<1%
ptv_ns_b: subsets of missense variants with MAF<1%, predicted as deleterious by at least one of the five algorithms (Five algorithm: Polyphen2-HumDiv, PolyPhen2-HumVar, LRT, Mutation Taster, and SIFT)
ptv_ns_s: Identify subsets of missense variants with MAF<1%, predicted as deleterious by all five algorithms

Including both common and rare variants, we considered the following eight combinations of variants:

- COM (Single-SNP analysis)
- COM-GWAS (GWAS catalog + Single-SNP analysis)
- RARE1 (collapsed rare variants by CMC method)
- RARE2 (collapsed rare variants by weighted CMC method)

- ALL1 (Single-SNP analysis + collapsed rare variants by CMC method)
- ALL2 (Single-SNP analysis + collapsed rare variants by weighted CMC method)
- ALL1-GWAS (GWAS catalog + Single-SNP analysis + collapsed rare variants by CMC method)
- ALL2-GWAS (GWAS catalog + Single-SNP analysis + collapsed rare variants by weighted CMC method)

A.2. Variant selection

In the T2D data, we selected SNPs using 5-fold CV on the training set. In this case, we used SLR, LASSO and EN to select SNPs. The phenotype y_i of subject $i = 1, \dots, n$ was set as a dependent variable (T2D = 1, normal = 0), and the genotype x_{ij} of the j -th SNP ($j = 1, \dots, p$) for subject i as an independent variables with an additive genetic model (AA = 0, Aa = 1, aa=2, where A and a indicate the major and minor allele, respectively). Let $\pi_i = P(y_i = 1 | x)$. The SLR model is given by

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \gamma_1 \text{sex}_i + \gamma_2 \text{BMI}_i,$$

where β_0 and β_j 's are the intercept and effect sizes of SNPs, respectively. γ_1 and γ_2 represent the sex, BMI of the i -th individual respectively. Variable selection was performed by the Akaike's Information Criterion (AIC)-based stepwise procedure using R package, named "MASS" [23].

The LASSO and EN estimates of β were obtained by minimizing

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip} - \gamma_1 \text{sex}_i - \gamma_2 \text{BMI}_i)^2 + \lambda_1 \sum_{j=1}^p |\beta_j|$$

and

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip} - \gamma_1 \text{sex}_i - \gamma_2 \text{BMI}_i)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$$

respectively. The tuning parameters λ_1 and λ_2 are estimated using 10-fold CV. The penalized methods performed using R package, named "glmnet" [24].

A.3. Model prediction

To build a risk prediction model, we considered 7 prediction models. The four models were constructed by using SLR, LASSO, EN and SVM and the other three models selected the variables with SLR

and EN, and then constructed by using SVM (i.e., SLR-SVM, LASSO-SVM and EN-SVM). By combining eight variants combinations, we considered 56 combination of variable selection and prediction models. To compare the performance of these prediction models, we calculated the AUC by applying each risk prediction models using the test set.

B.1 Combination of the variants

We selected common variants and rare variants in the same way as WES data analysis using model building set. In addition, rare variants were selected using SKAT. The rare variant sets consisted of ptv_ms set. The number of SNPs was selected from 1 to 30 and 100 based on the P-value. Including both common and rare variants, we considered the following eight combinations of variants:

- COM (Single-SNP analysis)
- COM-GWAS (GWAS catalog and Single-SNP analysis)
- RARE-CMC (Collapsed rare variants by CMC method)
- RARE-WCMC (Collapsed rare variants by WCMC method)

- SKAT-CMC (Select rare variants using SKAT and collapsed rare variants by CMC method)
- SKAT -WCMC (Select rare variants using SKAT and collapsed rare variants by WCMC method)
- ALL-CMC (Single-SNP analysis and collapsed rare variants by CMC method)
- ALL-WCMC (Single-SNP analysis and collapsed rare variants by WCMC method)
- ALL-SKAT-CMC (Single-SNP analysis, select rare variants using SKAT and collapsed rare variants by CMC method)
- ALL-SKAT-WCMC (Single-SNP analysis, select rare variants using SKAT and collapsed rare variants by WCMC method)
- ALL-CMC-GWAS (GWAS catalog, Single-SNP analysis and collapsed rare variants by CMC method)
- ALL-WCMC-GWAS (GWAS catalog, Single-SNP analysis and collapsed rare variants by weighted CMC method)

- ALL-SKAT-CMC-GWAS (GWAS catalog, Single-SNP analysis, select rare variants using SKAT and collapsed rare variants by CMC method)
- ALL-SKAT-WCMC-GWAS (GWAS catalog, Single-SNP analysis, select rare variants using SKAT and collapsed rare variants by weighted CMC method)

B.2 variants selection

The variants are selected using SLR, LASSO, and EN on the training set. In the case of SLR, the analysis was conducted based on AIC and AUC. We consisted a set of SNPs having non-zero coefficients in 5-fold CV for SLR, LASSO, EN.

B.3 Model building and prediction

We considered 13 combinations of variants selection and prediction methods (i.e., SLR-SLR, SLR-LASSO, SLR-EN, LASSO-SLR, LASSO-LASSO, LASSO-EN, EN-SLR, EN-LASSO, EN-EN, SLR-SVM, LASSO-SVM, EN-SVM, and SVM). For each combination, the risk prediction models were built using the model building set.

Chapter 4

Result

A. WES data analysis

The genetic association with T2D was analyzed using logistic regression with adjustment for sex and BMI as covariates for each variants combinations. For the common variant analysis, the 39 common variants were selected from data based on p value less than 1.0×10^{-3} . In addition, the 24 common variants were found to be associated with T2D in GWAS catalog. For the rare variant analysis, Table 3 shows the number of genes (collapsed rare variants) with p-values less than 1.0×10^{-2} by the CMC method categorized by rare variant sets.

Table 3 The number of genes selected by each variant list and collapsed method

<i>Rare variant sets</i>	<i>CMC</i>	<i>WCMC</i>
ptv	17	14
ptv_ms	96	100
ptv_ns_b	16	15
ptv_ns_s	15	14

Table 4 shows the prediction results showing the AUC values for the test data sets. First, the case when using the SVM prediction model showed the largest AUC value of 0.8961 in ptv set with ALL-WCMC group. Second, the case when using the SVM prediction model with LASSO selected variables generated the largest AUC value of 0.9291 in ptv_ms set with ALL-CMC group. Third, the case when using the SVM prediction model showed the largest AUC value of 0.8924 in ptv_ns_b set with ALL-WCMC group. Lastly, the case when using the SVM prediction model showed the largest AUC value of 0.8961 in ptv_ns_s set with ALL-WCMC group. The case when using the SVM prediction model with LASSO selected variables yielded the largest AUC value of 0.9291 in all group.

Table 4 AUC of the test data set

Rare variant sets	variant groups	LASSO	ES	SLR	SVM	LASSO-SVM	EN-SVM	SLR -SVM
-	COM	0.8258	0.8243	0.839	0.8627	0.8609	0.8623	0.8634
	COM-GWAS	0.821	0.8195	0.8371	0.8482	0.8563	0.855	0.8594
Ptv	RARE1	0.7389	0.7398	0.7567	0.773	0.773	0.7739	0.7739
	RARE2	0.7412	0.7359	0.7551	0.7742	0.7723	0.773	0.7728
	ALL1	0.8415	0.841	0.8517	0.8856	0.8875	0.8868	0.8896
	ALL2	0.8382	0.8362	0.8517	0.8961	0.8771	0.8894	0.895
	ALL1-GWAS	0.8386	0.8344	0.8481	0.8748	0.8751	0.8758	0.8817
	ALL2-GWAS	0.8345	0.8288	0.8507	0.883	0.8639	0.883	0.891
ptv_ms	RARE1	0.8901	0.8902	0.881	0.865	0.8684	0.8684	0.8655
	RARE2	0.889	0.8883	0.8937	0.8913	0.8907	0.8915	0.8803
	ALL1	0.916	0.9168	0.9222	0.8927	0.8938	0.8937	0.9
	ALL2	0.9158	0.8763	0.9253	0.9152	0.9275	0.9152	0.9194
	ALL1-GWAS	0.9121	0.9121	0.9182	0.8944	0.895	0.891	0.8949
	ALL2-GWAS	0.9089	0.8664	0.9236	0.9214	0.9291	0.9214	0.9215
ptv_ns_b	RARE1	0.7358	0.7368	0.7554	0.7572	0.7554	0.7562	0.7634
	RARE2	0.7368	0.7346	0.7557	0.7579	0.7496	0.7539	0.7611
	ALL1	0.8425	0.843	0.8512	0.8896	0.8861	0.8843	0.8889
	ALL2	0.8458	0.8431	0.8512	0.8924	0.8893	0.8877	0.8905
	ALL1-GWAS	0.8386	0.8388	0.8473	0.8737	0.8735	0.8741	0.8782
	ALL2-GWAS	0.8398	0.833	0.8483	0.887	0.8732	0.8769	0.8894

ptv_ns_s	RARE1	0.7351	0.7372	0.7567	0.773	0.7719	0.7709	0.7739
	RARE2	0.7349	0.7341	0.7551	0.7742	0.7675	0.7702	0.7728
	ALL1	0.8421	0.8427	0.8517	0.8856	0.8866	0.8869	0.8896
	ALL2	0.8451	0.8373	0.8517	0.8961	0.8808	0.8755	0.895
	ALL1-GWAS	0.8368	0.8379	0.8481	0.8748	0.8749	0.8777	0.8817
	ALL2-GWAS	0.8375	0.8319	0.8507	0.883	0.8652	0.8638	0.891

Table 4 is summarized in Figure 2 to Figure 5. As shown in the figures, prediction performance using both common variants and rare variants is better than using only common variants or only rare variants. The ptv_ms set yielded the highest AUC value among the four rare variant sets.

Figure 2 Result of each variant combination and each method using ptv set

Figure 3 Result of each variant combination and each method using ptv_ms set

Figure 4 Result of each variant combination and each method using ptv_ns_b set

Figure 5 Result of each variant combination and each method using ptv_ns_s set

B. WES and array data analysis

Initially, as the number of variables increases, the AUC value gradually increases. In the case of stepwise based on AIC, the largest AUC was observed when the number of variants was 11. The AUC of 0.7134 was the largest when selecting variants with SLR and modelling with SLR with ALL_SKAT_CMC group. In the case of stepwise based on AUC, the largest AUC was observed when the number of variants was 13. The AUC of 0.7222 was the largest when selecting variants with SLR and modelling with SLR with ALL_SKAT_CMC group. Thereafter, as the number of variables increases, the AUC value gradually decreases. Fig 4 shows the results when the number of variants was from 10 to 14. The AUC value of the risk prediction model by logistic regression using only covariates (gender and BMI) is 0.6758, which is the black line in Fig 6.

In summary, it is better to use both common and rare variants than to use only common variants or rare variations. The performance of the risk prediction model was best when using the ptv_ms set among the four rare variant sets.

(a) SLR based on AIC

(b) SLR based on AUC

Figure 6 Results for KAR and T2D-GENES data. Each bar represents one of fourteen SNP data sets

Chapter 5

Discussion

In this study, we used statistical methods (SLR, LASSO, EN and SVM) to build risk prediction models. Then, we compared the performance of the risk prediction models by each variant group (COM, COM-GWAS, RARE-CMC, RARE-WCMC, ALL-CMC, ALL-WCMC, ALL-CMC-GWAS, ALL-WCMC-GWAS). As a result, the performance of the risk prediction models using `ptv_ms` rare variant set was better than that of the prediction models using other rare variant sets. In the model building, SVM showed better performance than other methods. The performance of the risk prediction model is better than not using the SNPs reported in the GWAS catalog. As a result of other cases, the AUC value of the training set continued to increase as the number of variables increased. However, the AUC value of the test set tended to increase at first and then decrease. In the most cases of SLR,

the performance of risk prediction models based on the AUC is better than the performance of risk prediction models based on the AIC.

In addition, we will compare performance by applying the model to other data set. We also perform analysis with other continuous traits such as BMI and metabolic syndrome related traits.

Bibliography

- 1 Kooperberg, C., M. LeBlanc, and V. Obenchain, Risk Prediction Using Genome-Wide Association Studies. *Genetic Epidemiology*, 2010. 34(7): p. 643-652.
- 2 Wang, W.Y.S., et al., Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, 2005. 6(2): p. 109-118.
- 3 Hoerl, A.E. and R.W. Kennard, Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1970. 12(1): p. 55-&.
- 4 Hoerl, A.E. and R.W. Kennard, Ridge Regression - Applications to Nonorthogonal Problems. *Technometrics*, 1970. 12(1): p. 69-&.
- 5 Hoerl, A.E., Ridge Regression. *Biometrics*, 1970. 26(3): p. 603-&.
- 6 Tibshirani, R., Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, 1996. 58(1): p. 267-288.
- 7 Zou, H. and T. Hastie, Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 2005. 67: p. 301-320.
- 8 Wei, Z., et al., Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease. *American Journal of Human Genetics*, 2013. 92(6): p. 1008-1012.
- 9 Burges, C.J.C., A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998. 2(2): p. 121-167.
- 10 Cortes, C. and V. Vapnik, Support-Vector Networks. *Machine Learning*, 1995. 20(3): p. 273-297.
- 11 Yoon, D., Y.J. Kim, and T. Park, Phenotype prediction from genome-wide association studies: application to smoking behaviors. *Bmc Systems Biology*, 2012. 6.
- 12 Wei, Z., et al., From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *Plos Genetics*, 2009. 5(10).
- 13 Manolio, T.A., et al., Finding the missing heritability of complex diseases. *Nature*, 2009. 461(7265): p. 747-753.
- 14 Gibson, G., Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 2012. 13(2): p. 135-145.
- 15 Goldstein, J.L. and M.S. Brown, Ldl Receptor Locus and the Genetics of Familial Hypercholesterolemia. *Annual Review of Genetics*, 1979. 13: p. 259-289.
- 16 Cirulli, E.T. and D.B. Goldstein, Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 2010. 11(6): p. 415-425.
- 17 Stankiewicz, P. and J.R. Lupski, Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, 2010. 61: p. 437-455.
- 18 Tennessen, J.A., et al., Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, 2012. 337(6090): p. 64-69.
- 19 Choi, S., S. Bae, and T. Park, Risk Prediction Using Genome-Wide Association Studies on Type 2 Diabetes. *Genomics & informatics*, 2016.

- 14(4): p. 138-148.
- 20 Fuchsberger, C., et al., The genetic architecture of type 2 diabetes. *Nature*, 2016. 536(7614): p. 41-+.
 - 21 Welter, D., et al., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 2014. 42(D1): p. D1001-D1006.
 - 22 Li, B.S. and S.M. Leal, Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, 2008. 83(3): p. 311-321.
 - 23 Ripley, B., et al., Package 'MASS'. CRAN Repos. [Httpcran R-Proj. Org web packages MASS MASS Pdf](http://cran.r-project.org/web/packages/MASS/MASS.pdf), 2013.
 - 24 Friedman, J., T. Hastie, and R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 2010. 33(1): p. 1-22.

초 록

전장유전체 연관분석연구 (GWAS, genome-wide association study)를 통하여 질병과 관련된 많은 대립유전자 빈도가 비교적 큰 변이 (common variant)를 많이 찾아내었다. 몇몇 연구에서는 이러한 대립유전자 빈도가 비교적 큰 변이를 이용하여, 벌점화 회귀분석 (penalized regression) 또는 통계학습방법 (statistical learning methods)을 기초로 한 예측모형을 제안하였다. 그러나 대립유전자 빈도가 비교적 큰 변이 (common variant)만으로 표현형을 설명하는데 충분하지 않다. 이러한 문제점을 해결하기 위한 한 가지 방법으로 빈도가 낮은 변이 (rare variant)를 고려할 수 있다. 그 이유는 빈도가 낮은 변이들은 질병에 대해 영향이 크기 때문이다. 최근 차세대 염기서열 분석 (NGS, Next Generation Sequencing) 기술의 발달로 인해 질병과 관련된 빈도가 낮은 변이들이 발견되었다. 그러나 단지 몇몇 연구에서만 빈도가 높은 변이와 빈도가 낮은 변이를 이용하여 만든 예측모형의 비교를 하였다. 이 연구에서는 T2D-GENES 컨소시엄의 전체 엑솜 염기서열 분석 (WES, whole-exome sequencing) 자료의 빈도가 높은 변이와 빈도가 낮은 변이와 한국인 유전체 분석사업 (KARE) 자료의 빈도가 높은 변이를 이용하여 예측모형을 만들고 체계적인 성능비교를 하였다. 먼저, 단계적 회귀분석, LASSO, Elastic-Net 그리고 지지 벡터 머신 (SVM, support vector machine)을 이용하여 예측모형을 만들었다. 그 후 예측모형의 ROC 커브의 밀면적 (AUC, Area under a ROC curve)을 구하여 비교하였다. 결과에서 보면 빈도가 높은 변이만 이용하여 예측모형을 만든 경우와 빈도가 낮은 변이만 이용하여 예측모형을 만든 경우보다 빈도가 높은 변이와 빈도가 낮은 변이를

같이 사용하여 예측모형을 만들었을 때 성능이 더 좋았다. 예측모형을 만들 때 사용한 변이 구성에 따라 ROC 커브의 밑면적 값이 달랐으나, 다른 방법들에 비해 지지 벡터 머신으로 모형을 만들었을 때 ROC 커브의 밑면적 값이 컸다. 또한, 4 가지 빈도가 낮은 변이 구성중 pTv_ns 구성일 때 ROC 커브의 밑면적이 크게 나타났다.

주요어: 전체 엑솜 염기서열 분석, 위험 예측 모형, 제 2 형 당뇨병, 별점화 회귀분석, 단계적 회귀분석, 지지 벡터 머신

학 번: 2015-20506