Risk prediction of type 2 diabetes using common and rare variants

Sunghwan Bae

Interdisciplinary Program in Bioinformatics,
Seoul National University,
Seoul, South Korea
Email: beash1@snu.ac.kr

Taesung Park*

Department of Statistics,
Seoul National University,
Seoul, South Korea
Email: tspark@stats.snu.ac.kr
*Corresponding author

Abstract: The recent development of next generation sequencing technology has led to the identification of several disease-related genetic variants. In this study, we systematically compare the performance of prediction models using common and rare variants from the Whole Exome Sequencing data of the Type 2 Diabetes Genetic Exploration by Next generation sequencing in multi-ethnic samples. We evaluated several methods for predicting binary phenotypes such as Stepwise Logistic Regression, Penalised Regression and Support Vector Machine (SVM). We first constructed prediction models by combining variable selection and prediction methods for Type 2 Diabetes. We then calculated the Area Under the Curve (AUC) to compare the performance of the prediction models. The results indicate that the performance of the common and rare variants combination was better than either that of the common variants only or the rare variants only. Further, the AUC values of SVM were always larger than those of other prediction models.

Keywords: WES; whole exome sequencing; risk prediction model; T2D; type 2 diabetes; penalised regression methods; stepwise selection; SVM; support vector machine.

Reference to this paper should be made as follows: Bae, S. and Park, T. (2018) 'Risk prediction of type 2 diabetes using common and rare variants', *Int. J. Data Mining and Bioinformatics*, Vol. 20, No. 1, pp.77–90.

Biographical notes: Sunghwan Bae received his BSc in molecular biology from Pusan National University, Korea, in 2014. He is a master`s student in bioinformatics at Seoul National University, Korea.

Taesung Park received his PhD degree in Biostatistics from the University of Michigan, USA, in 1990. He worked as an associate professor from September 1999 to September 2001 in the Department of Statistics at Seoul National University, Korea, where he has been a professor since October 2001. His research areas include genome-wide association studies and statistical genetics.

This paper is a revised and expanded version of a paper entitled 'Risk prediction using common and rare genetic variants: application to Type 2 diabetes' presented at the 'IEEE International Conference on Bioinformatics and Biomedicine', Kansas City, MO, USA, 13–16 November 2017.

1 Introduction

Genome-Wide Association Studies (GWAS) have identified many common genetic variants associated with diseases. Several studies using these variants to predict disease have achieved improved risk prediction (Kooperberg et al., 2010). For example, several studies have shown the usefulness of Multiple Logistic Regression (MLR), a traditional approach, in predicting disease risk (Lindstrom et al., 2012; Jostins and Barrett, 2011; Wacholder et al., 2010). In recent years, some direct-to-consumer (DTC) companies, such as 23andMe (http://www.23andme.com) and Pathway Genomics (https://www.pathway.com), offer personal genomic information services. However, there are some limitations to disease risk prediction using genetic variants. The first difficulty in constructing a disease risk prediction model lies in the 'large p, small n problem'. That is, the number of genetic variants exceeds the number of individuals. Second, the Linkage Disequilibrium (LD) between genetic variants reduces the statistical power for confirming significant associations (Wang et al., 2005). The high variance of the coefficient estimates is caused by the multi-collinearity due to the high LD among the genetic variants. Finally, one must consider the small effect size of genetic variants on disease as well as missing heritability (Manolio et al., 2009).

Recently, various statistical approaches have been developed to solve this problem. Penalised regression approaches such as ridge (Hoerl and Kennard, 1970b; Hoerl and Kennard, 1970a; Hoerl, 1970), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), and Elastic-Net (Zou and Hastie, 2005) have been proposed to solve 'large p, small n' problems. The penalised method approach in high-dimensional data has shown better results than the non-penalised method. For example, the accuracy of risk prediction of Crohn's and inflammatory bowel diseases was improved by the utilisation of a large number of Single Nucleotide Polymorphisms (SNPs) with penalised regression approaches (Wei et al., 2013; Kooperberg et al., 2010). Data mining approaches have also been widely applied to improve risk prediction performance. Specifically, Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Burges, 1998) was shown to have better performance than other classification algorithms (Yoon et al., 2012), particularly when using multiple SNPs (Wei et al., 2009).

The poor performance of many risk prediction studies using common variants could be due to the small effect size of the common variants (Kraft and Hunter, 2009). On the other hand, the development of genotyping techniques such as Next-Generation Sequencing (NGS) has also identified rare genetic variants associated with diseases. Furthermore, recent studies have shown that rare variants play a key role in complex diseases with large effect sizes such as autism, atherosclerosis and mental retardation (Cirulli and Goldstein, 2010; Gibson, 2012; Goldstein and Brown, 1979; Stankiewicz and Lupski, 2010). From an evolutionary point of view, rare variants are more harmful than common variants (Tennessen et al., 2012). Our earlier work compared the performance of risk prediction models using common variants (Choi et al., 2016). We first built the prediction model by combining variant selection and prediction methods for 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 study, we compare risk prediction models using both common and rare variants with those using only one of the variant types from the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D GENES) consortium data. We made three different risk prediction models using common variants only, rare variants only, and both common and rare variants, then compared their performance. We further compared the performance of the risk prediction models by calculating Area Under the Curve (AUC) values.

Specifically, we performed two comparison analyses. The first analysis utilised Whole Exome Sequencing (WES) data of East Asian populations from the T2D-GENES consortium. 2,154 Korean and Chinese samples were used. The second analysis used only the Korean dataset, 1,087 WES Korean samples from the T2D-GENES consortium along with additional SNP chip data. Hereafter, we refer to the first analysis as 'WES analysis' and the second one as 'WES+SNP chip analysis'.

2 Materials

The T2D-GENES consortium is a large-scale collaborative study aimed at discovering genetic variants affecting the risk of T2D. The 12,940 individuals (6504 with T2D and 6436 controls) were collected from five ancestry groups (4541 Europeans and around 2000 each of East Asians, South Asians, American-Hispanics, and African-Americans). The samples were genotyped using WES (Loh et al., 2016). For our first WES analysis, we used 2154 East Asian samples of Korean and Chinese in Singapore from T2D-GENES consortium data. Table 1 summarises the demographic information of WES samples.

Table 1 Demographic variables for Korea and Chinese from T2D-GENES

Information	*Korea*		*Chinese*	
	T2D	*Non-T2D*	*T2D*	*Non-T2D*
# of sample	526	561	486	592
Male/Female	286/240	232/328	233/253	229/363
Age (year)	53.82 ± 3.58	63.25 ± 7.47	58.04 ± 9.33	58.25 ± 7.02
Body Mass Index (BMI)	25.69 ± 3.25	23.72 ± 3.06	25.60 ± 3.80	22.85 ± 3.38

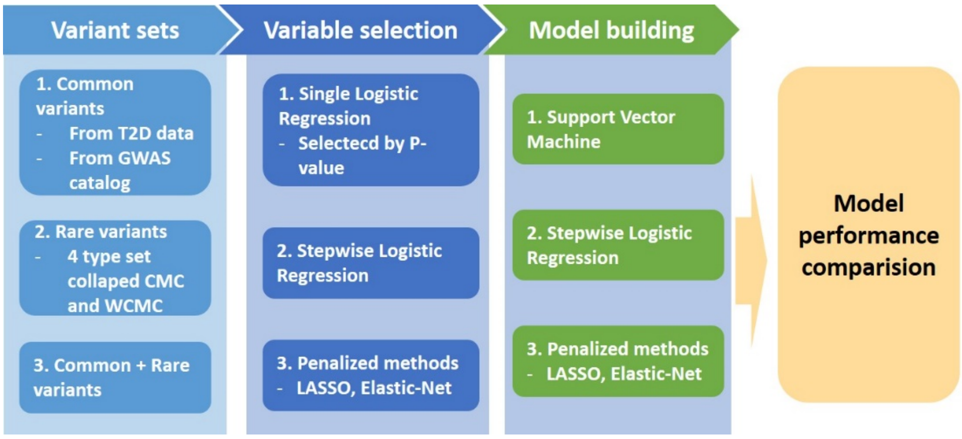
For the WES+SNP analysis, we used only the 1087 Korean samples included in the T2D-GENES consortium, which were selected from the 8838-sample database of the Korea Association Resource (KARE) project. The KARE project is a regional society-based cohort from the Ansung and Ansan recruiting areas. Its 8838 samples were genotyped using Affymetrix Genome-Wide Human SNP Array 5.0. Missing genotypes were imputed using the Eagle program v2.3.5 (Loh et al., 2016). For the WES+SNP chip analysis, we used 1087 Korean samples containing both WES and SNP chip data.

3 Methodology

3.1 WES analysis

This analysis was performed using the WES data from the T2D consortium using 5-fold cross validation (CV). The analysis was conducted in the same way for the Korean and Chinese datasets. We selected the common variants using single-SNP analysis and the GWAS catalogue (Welter et al., 2014). We selected the rare variants using the Combined Multivariate and Collapsing (CMC) method (Li and Leal, 2008), Weighted CMC (WCMC) method and Sequence Kernel Association test (SKAT) (Wu et al., 2011). We randomly split the data into 5 CV, then made a training set using 4 CV and a test set using the other 1 CV. At Step 1, the variants were selected by SLR, Logistic Regression LASSO (LR-LASSO), and Logistic Regression EN (LR-EN) using the training set. At Step 2, the risk prediction models were built by SLR, LR-LASSO, LR-EN, and SVM using the training set. At step 3, the AUC values were calculated using the test set. Figure 1 summarises the analysis scheme.

Figure 1 The overall analysis scheme of model building



3.1.1 Combination of the variants

The common variants were collected from the GWAS catalogue for T2D and selected by single-SNP analysis with adjustment for gender and Body Mass Index (BMI). The SNPs were selected based on the P-value ($< 10 \times 10^{-3}$). We first collapsed rare variants based on the CMC and WCMC methods. We then fit a gene-based logistic regression model with adjustment for gender and BMI. The genes were selected based on the P-value ($< 10 \times 10^{-2}$). The rare variant sets consisted of four types based on functional annotation and Minor Allele Frequency (MAF) < 1%. We used annotations from CHAoS v0.6.3, SnpEFF v3.1, and VEPv2.7. The ptv set was defined as a protein-truncating (for example, nonsense, frameshift, essential splice site) variants set. The ptv_ms set was defined as a protein-altering (missense, inframeshift, non-essential splice site) variants set. Subsets of missense variants were identified using annotation predictions from Polyphen2-HumDiv, PolyPhen2-HumVar, LRT, Mutation Taster and SIFT. The ptv_ns_b set was defined as those variants predicted to be deleterious by at least one algorithm. The ptv_ns_s set was defined as those variants predicted to be deleterious by all five algorithms. Table 2 summarises the eight combinations of variants.

Table 2 Eight combinations of variants in the WES data analysis

COM: Single-SNP analysis
COM-GWAS: GWAS catalogue and Single-SNP analysis
RARE-CMC: collapsed rare variants by CMC method
RARE-WCMC: 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-CMC-GWAS: GWAS catalogue, Single-SNP analysis and collapsed rare variants by CMC method
ALL-WCMC-GWAS: GWAS catalogue, Single-SNP analysis and collapsed rare variants by weighted CMC method

3.1.2 Variant selection

The variants were selected using SLR, LR-LASSO, and LR-EN on the training set. The dependent variables (T2D = 1, non-T2D = 0) are y_i of subject i = 1, …, n and the independent variables are genotype x_{ij} of j-th SNP for subject i with an additive genetic model. Let $\pi_i = P(y_i = 1 \mid x)$. The intercept and coefficients of SNPs are denoted by β_0 and β_j's, respectively. The coefficients of gender and BMI are denoted by γ_1 and γ_2, respectively. The SLR with adjustment for gender and BMI model is given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \gamma_1\, gender_i + \gamma_2\, BMI_i.$$

The SLR was performed based on Akaike's Information Criterion (AIC) using R-packages *MASS* (Ripley et al., 2013).

The LR-LASSO and LR-EN solve the following equations:

$$\min_{\beta_0,\beta,\gamma}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij} - \gamma_1 gender_i - \gamma_2 BMI_i\right)^2 + P_\lambda(\beta)\right].$$

LR-LASSO penalty is defined as $P_\lambda(\beta) = \lambda\sum|\beta|$ and LR-EN penalty is defined $P_\lambda(\beta) = \lambda\left[(1-\alpha)\sum|\beta| + \alpha\sum\beta^2\right]$. The tuning parameters are λ and α. We performed LR-LASSO and LR-EN using the R-package *glmnet* (Friedman et al., 2010).

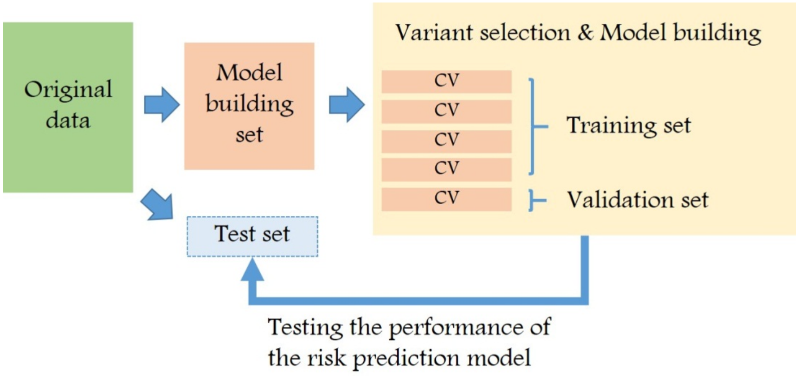
3.1.3 Model building and prediction

We considered seven risk prediction models. The four models were constructed using SLR, LR-LASSO, LR-EN and SVM. The other three models with the variables selected by SLR, LR-LASSO and LR-EN were constructed by using SVM. To denote the combinations of variants selection and prediction methods we use 'A'/'a', where 'A' represents the method of variant selection and 'a' indicates the prediction model. For example, three models with the variables selected by SLR, LR-LASSO and LR-EN constructed by SVM are denoted by SLR/SVM, LR-LASSO/SVM, and LR-EN/SV, respectively. Thus, from eight variant sets and seven prediction models, we analysed 56 combinations of variant selection and risk prediction models. The AUC values were calculated by applying each risk prediction model to the test set.

3.2 WES + SNP chip analysis

We performed this analysis by combining the SNPs of the same individuals in both T2D-GENES consortium and KARE data. One-fifth of the data was then separated into test sets and the risk prediction models were created using the remaining data in the same way via the 5-fold CV. After that, AUC values were calculated by applying the models to the test set. Figure 2 summarises the analysis scheme.

Figure 2 The overall analysis scheme of WES and SNP chip analysis


Original data
Model building set
Variant selection & Model building
CV
CV
CV
CV
CV
Training set
Validation set
Test set
Testing the performance of the risk prediction model

3.2.1 Combination of the variants

We selected common and rare variants in the same way as in the WES data analysis, using a model-building set. In addition, rare variants were selected using SKAT. The rare variant sets comprised the ptv_ms set. We selected the 30 highest-ranking SNPs based on P-values as well as the top 100 SNPs. Table 3 summarises 14 combinations of variants.

3.2.2 Variant selection

The variants are selected using SLR, LR-LASSO, and LR-EN on the training set. In the case of SLR, the analysis was conducted based on AIC and AUC (Kim et al., 2015). We consisted a set of SNPs having non-zero coefficients in 5-fold CV for SLR, LR-LASSO, LR-EN.

3.2.3 Model building and prediction

We considered 13 combinations of variant selection and prediction methods. The variants were selected by SLR, LR-LASSO and LR-EN, and then the models were constructed using SLR, LR-LASSO, LR-EN and SVM. Thus, the following 12 combinations were used: SLR/SLR, SLR/LR-LASSO, SLR/LR-EN, LR-LASSO/SLR, LR-LASSO/LR-LASSO, LR-LASSO/LR-EN, LR-EN/SLR, LR-EN/LR-LASSO, LR-EN/LR-EN, SLR/SVM, LR-LASSO/SVM and LR-EN/SVM. The last model was constructed using SVM without variant selection. For each combination, the risk prediction models were built using the model building set.

Table 3 WES and SNP chip analysis

COM: Single-SNP analysis

COM-GWAS: GWAS catalogue 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 catalogue, Single-SNP analysis and collapsed rare variants by CMC method

ALL-WCMC-GWAS: GWAS catalogue, Single-SNP analysis and collapsed rare variants by weighted CMC method

ALL-SKAT-CMC-GWAS: GWAS catalogue, Single-SNP analysis, select rare variants using SKAT and collapsed rare variants by CMC method

ALL-SKAT-WCMC-GWAS: GWAS catalogue, Single-SNP analysis, select rare variants using SKAT and collapsed rare variants by weighted CMC method

4 Result

4.1 WES analysis

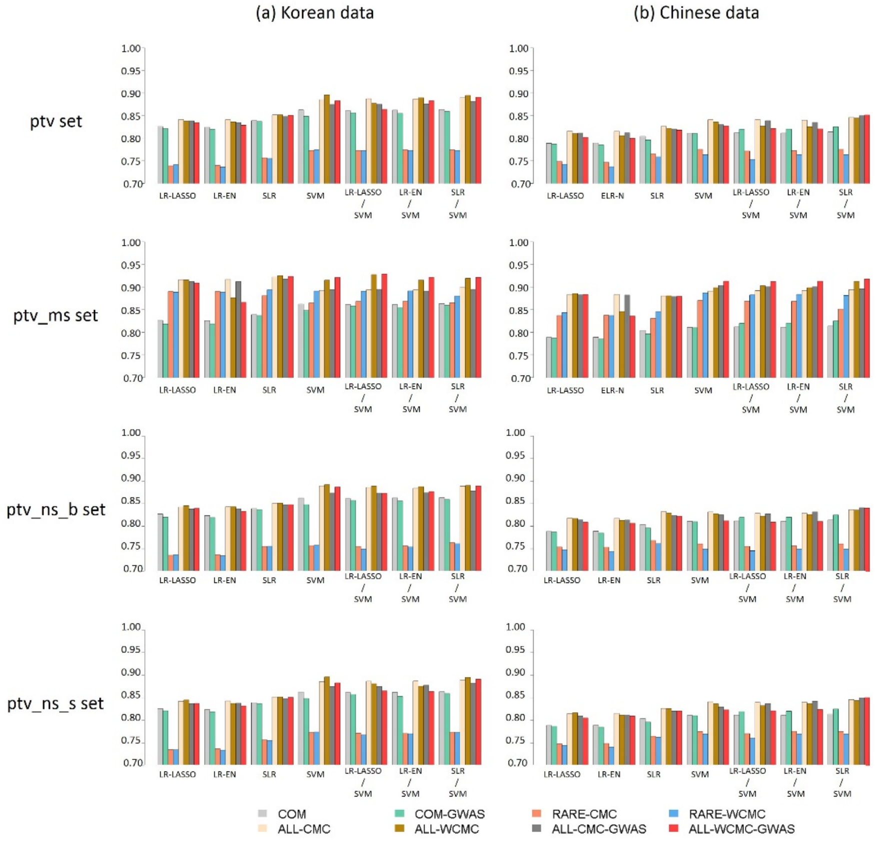
The genetic association with T2D was analysed using logistic regression with adjustment for gender and BMI as covariates for each variant combination for the Korean and Chinese datasets, respectively. For the common variant analysis, the 39 common variants were selected from the Korean data and the 24 common variants from the Chinese data based on a P-value less than 1.0×10^{-3}. In addition, the 24 common variants were found to be associated with T2D in the GWAS catalogue. Regarding the rare variant analysis, Table 4 shows the number of genes (collapsed rare variants) with P-values less than 1.0×10^{-2} based on the CMC method categorised by rare variant sets.

Figure 3 shows the prediction results for the Korean and Chinese data. The AUC was the largest when using the ptv_ms set among the four rare variant sets. In the T2D-GENES Korean data, the AUC of 0.9291 was thelargest when selecting variants with LR-LASSO and modelling with SVM in the ptv_ms set. In the Chinese data, the AUC wasthe largest when selecting variables with SLR and modelling with SVM in all 32 prediction models from four rare variant sets and eight variant groups. For the ptv_ms rare variant set, the ALL-WCMC-GWAS group has the largest AUC value, 0.9179.

Table 4 The number of genes selected by each variant list and collapsed method

Rare variant sets	*Korea*		*Chinese*	
	CMC	*Weighted CMC*	*CMC*	*Weighted CMC*
ptv	17	14	11	9
ptv_ms	96	100	61	60
ptv_ns_b	16	15	13	12
ptv_ns_s	15	14	10	11

Figure 3 Results for Korean and Chinese data. Each bar represents one of eight SNP data sets



4.2 WES+SNP chip analysis

Initially, as the number of variables increased, the AUC value gradually increased. 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 increased, the AUC value gradually decreased. Figure 4 shows the results when the number of variants increased from 10 to 14. The AUC value of the risk prediction model by logistic regression using only covariates (gender and BMI) was 0.6758, represented by the black line in Figure 4.

Figure 4 Results for KAR and T2D-GENES data. Each bar represents one of 14 SNP data sets

(a) SLR based on AIC

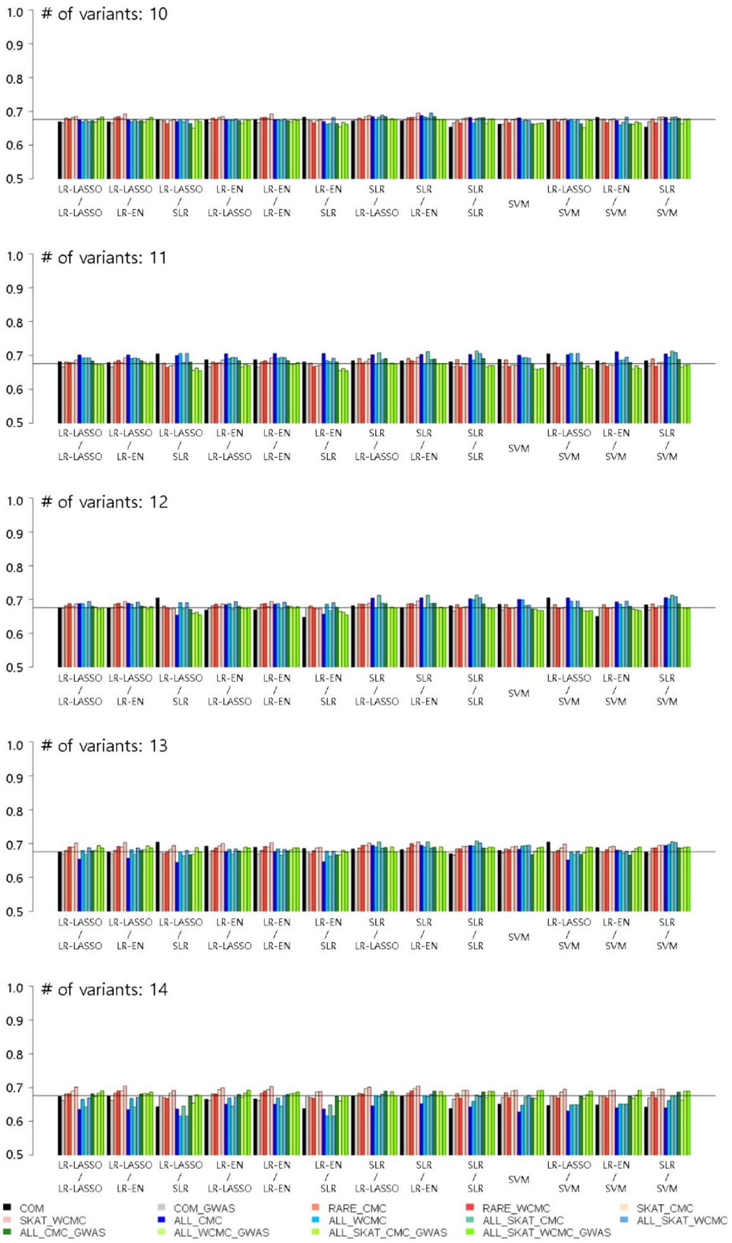
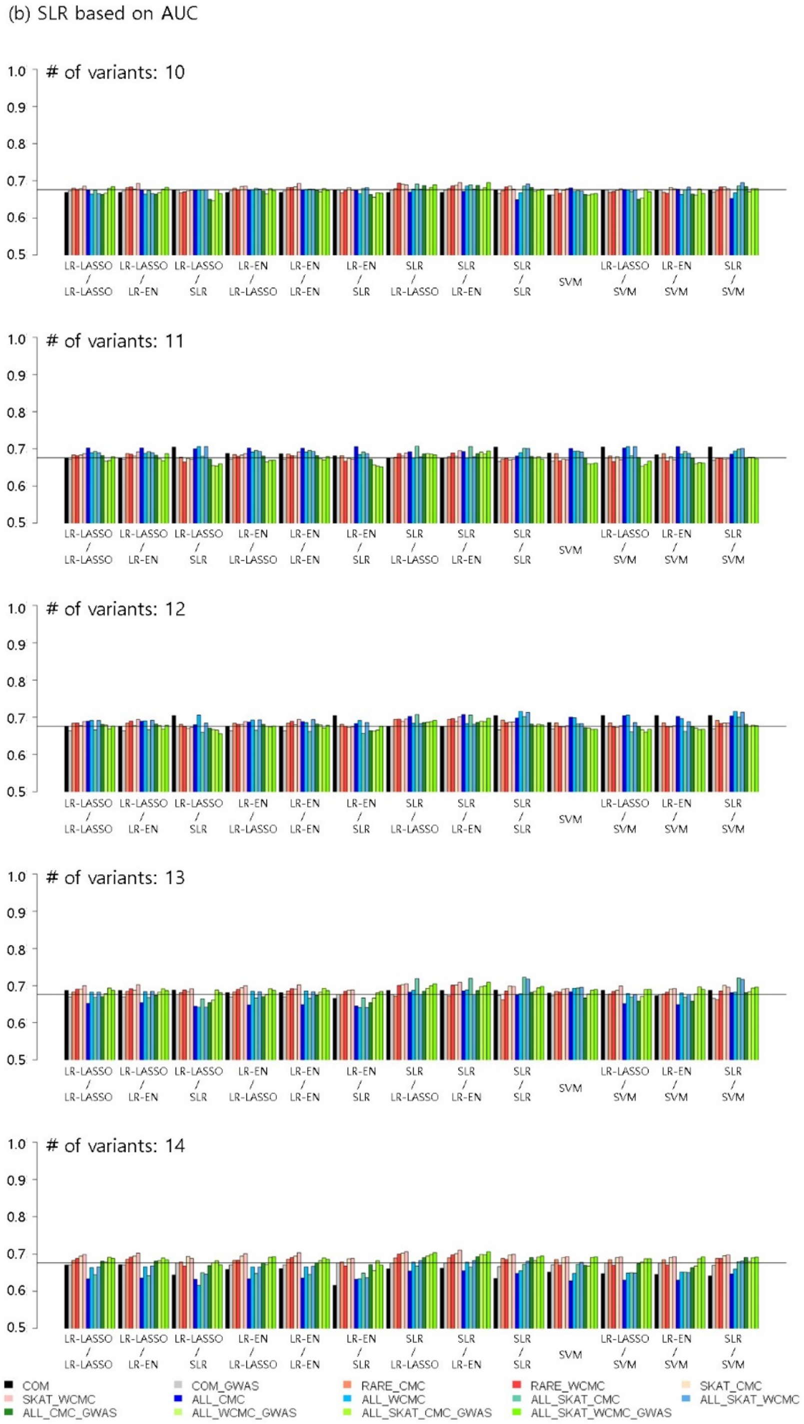


Figure 4 Results for KAR and T2D-GENES data. Each bar represents one of 14 SNP data sets (continued)



As shown in Figures 3 and 4, using both common and rare variants yields superior results to using only one of the variant types. The performance of the risk prediction model was best when using the ptv_ms set among the four rare variant sets.

5 Discussion

In this study, we compared the performance of three different risk prediction models: using common variants only, rare variants only, and both common and rare variants. We used several statistical methods such as SLR, LR-LASSO, LR-EN and SVM. We then performed two comparison analyses. The first WES analysis used 2,154 WES data of Korean and Chinese samples from the T2D-GENES consortium. The second WES+SNP chip analysis used 1,087 WES Korean samples from the T2D-GENES consortium and the SNP chip data. Overall, we have found that risk prediction models that include both non-genetic and genetic variants are more accurate than risk prediction models using only non-genetic variants. In both analyses, the performance of the model was best when both common and rare were used. Using risk prediction models, previous studies have reported some genes including rare variants such as AGXT2, PDZD8, EDN1, and IL10RB to be associated with T2D (Nepomuceno et al., 2017; Li et al., 2008; Anderssohn et al., 2014; Haddad, 2017).

In the WES analysis, the risk prediction model performed better when using the ptv_ms set than when using other rare variant sets. In the model building, SVM methods performed better than those of other methods. For the Korean data, the performance of risk prediction models not using SNPs reported in the GWAS catalogue was better than those of risk prediction models using SNPs reported in the catalogue. Conversely, for the Chinese data, risk prediction models using SNPs reported in the GWAS catalogue performed better than those of risk prediction models not using reported SNPs. The model performed poorly when constructed using one ancestry as training data and the other ancestry as test data. That is, when we built the prediction model using Korean data and applied it to the Chinese data, its performance was poor, and vice versa. This may be due to the different composition of common and rare variants between the Korean and Chinese samples.

The WES+SNP chip analysis showed that 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 later. In most cases of SLR, risk prediction models based on the AUC performed better than models based on the AIC. During the variable selection process, five SNP lists were generated from five CV. From the five SNP lists, we created five sets, from union to intersection (sets 1–4 consisting of SNPS present at least one, two, three, or four times in all five SNP lists, respectively, and set 5 consisting of all SNPs present in all five SNP lists). We made models using each SNP set and compared their performances. The performance of the risk prediction models was best when using the union of five SNP lists.

For further research, we plan to use other data to perform analyses with other continuous traits such as BMI and metabolic syndrome-related traits.

Acknowledgements

This research was supported by grants from 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, HI16C2037). This work was also supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

References

Anderssohn, M., Mclachlan, S., Luneburg, N., Robertson, C., Schwedhelm, E., Williamson, R.M., Strachan, M.W.J., Ajjan, R., Grant, P.J., Boger, R.H. and Price, J.F. (2014) 'Genetic and environmental determinants of dimethylarginines and association with cardiovascular disease in patients with type 2 diabetes', *Diabetes Care*, Vol. 37, pp.846–854.

Burges, C.J.C. (1998) 'A tutorial on Support Vector Machines for pattern recognition', *Data Mining and Knowledge Discovery*, Vol. 2, pp.121–167.

Choi, S., Bae, S. and Park, T. (2016) 'Risk prediction using genome-wide association studies on type 2 diabetes', *Genomics & informatics*, Vol. 14, pp.138–148.

Cirulli, E.T. and Goldstein, D.B. (2010) 'Uncovering the roles of rare variants in common disease through whole-genome sequencing', *Nature Reviews Genetics*, Vol. 11, pp.415–425.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, Vol. 20, pp.273–297.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software*, Vol. 33, pp.1–22.

Gibson, G. (2012) 'Rare and common variants: twenty arguments', *Nature Reviews Genetics*, Vol. 13, pp.135–145.

Goldstein, J.L. and Brown, M.S. (1979) 'Ldl receptor locus and the genetics of familial hypercholesterolemia', *Annual Review of Genetics*, Vol. 13, pp.259–289.

Haddad, S.A. (2017) *Gene-and pathway-based genomics of breast cancer and type 2 diabetes in African American women*, Boston University.

Hoerl, A.E. (1970) 'Ridge Regression', *Biometrics*, Vol. 26, p.603.

Hoerl, A.E. and Kennard, R.W. (1970a) Ridge Regression – applications to nonorthogonal problems', *Technometrics*, Vol. 12, p.69.

Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression – biased estimation for nonorthogonal problems', *Technometrics*, Vol. 12, p.55.

Jostins, L. and Barrett, J.C. (2011) 'Genetic risk prediction in complex disease', *Human Molecular Genetics*, Vol. 20, pp.R182–R188.

Kim, Y., Lee, S., Kwon, M.-S., Na, A., Choi, Y., Yi, S.G., Namkung, J., Han, S., Kang, M. and Kim, S.W. (2015) 'Developing cancer prediction model based on stepwise selection by AUC measure for proteomics data', *Bioinformatics and Biomedicine (BIBM), IEEE International Conference on*, IEEE, pp.1345–1350.

Kooperberg, C., Leblanc, M. and Obenchain, V. (2010) 'Risk prediction using genome-wide association studies', *Genetic Epidemiology*, Vol. 34, pp.643–652.

Kraft, P. and Hunter, D.J. (2009) 'Genetic risk prediction – are we there yet?' *New England Journal of Medicine*, Vol. 360, pp.1701–1703.

Li, B.S. and Leal, S.M. (2008) 'Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data', *American Journal of Human Genetics*, Vol. 83, pp.311–321.

Li, H.T., Louey, J.W.C., Choy, K.W., Liu, D.T.L., Chan, W.M., Chan, Y.M., Fung, N.S.K., Fan, B.J., Baum, L., Chan, J.C.N., Lam, D.S.C. and Pang, C.P. (2008) 'EDN1 Lys198Asn is associated with diabetic retinopathy in type 2 diabetes', *Molecular Vision*, Vol. 14, pp.1698–1704.

Lindstrom, S., Schumacher, F.R., Cox, D., Travis, R.C., Albanes, D., Allen, N.E., Andriole, G., Berndt, S.I., Boeing, H., Bueno-De-Mesquita, H.B., Crawford, E.D., Diver, W.R., Gaziano, J.M., Giles, G.G., Giovannucci, E., Gonzalez, C.A., Henderson, B., Hunter, D.J., Johansson, M., Kolonel, L.N., Ma, J., Le Marchand, L., Pala, V., Stampfer, M., Stram, D.O., Thun, M.J., Tjonneland, A., Trichopoulos, D., Virtamo, J., Weinstein, S.J., Willett, W.C., Yeager, M., Hayes, R.B., Severi, G., Haiman, C.A., Chanock, S.J. and Peter, K. (2012) 'Common genetic variants in prostate cancer risk prediction-results from the NCI breast and prostate cancer cohort consortium (BPC3)', *Cancer Epidemiology Biomarkers & Prevention*, Vol. 21, pp.437–444.

Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., Mccarthy, S., Abecasis, G.R., Durbin, R. and Price, A.L. (2016) 'Reference-based phasing using the Haplotype Reference Consortium panel', *Nature Genetics*, Vol. 48, pp.1443–1448.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., Mccarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., Mccarroll, S.A. and Visscher, P.M. (2009) 'Finding the missing heritability of complex diseases', *Nature*, Vol. 461, pp.747–753.

Nepomuceno, R., Villela, B.S., Corbi, S.C.T., Bastos, A.D., Dos Santos, R.A., Takahashi, C.S., Orrico, S.R.P. and Scarel-Caminaga, R.M. (2017) 'Dyslipidemia rather than type 2 diabetes mellitus or chronic periodontitis affects the systemic expression of pro-and anti-inflammatory genes', *Mediators of Inflammation*.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D. and Ripley, M.B. (2013) 'Package 'MASS'', *CRAN Repos.*

Stankiewicz, P. and Lupski, J.R. (2010) 'Structural variation in the human genome and its role in disease', *Annual Review of Medicine*, Vol. 61, pp.437–455.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W.Q., Kenny, E.E., Gravel, S., Mcgee, S., Do, R., Liu, X.M., Jun, G., Kang, H.M., Jordan, D., Leal, S.M., Gabriel, S., Rieder, M.J., Abecasis, G., Altshuler, D., Nickerson, D.A., Boerwinkle, E., Sunyaev, S., Bustamante, C.D., Bamshad, M.J., Akey, J.M., Go, B., Go, S. and Project, N.E.S. (2012) 'Evolution and functional impact of rare coding variation from deep sequencing of human exomes', *Science*, Vol. 337, pp.64–69.

Tibshirani, R. (1996) 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society Series B-Methodological*, Vol. 58, pp.267–288.

Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H.S., Diver, W.R., Thun, M.J., Cox, D.G., Hankinson, S.E., Kraft, P., Rosner, B., Berg, C.D., Brinton, L.A., Lissowska, J., Sherman, M.E., Chlebowski, R., Kooperberg, C., Jackson, R.D., Buckman, D.W., Hui, P., Pfeiffer, R., Jacobs, K.B., Thomas, G.D., Hoover, R.N., Gail, M.H., Chanock, S.J. and Hunter, D.J. (2010) 'Performance of common genetic variants in breast-cancer risk models', *New England Journal of Medicine*, Vol. 362, pp.986–993.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) 'Genome-wide association studies: theoretical and practical concerns', *Nature Reviews Genetics*, Vol. 6, pp.109–118.

Wei, Z., Wang, K., Qu, H.Q., Zhang, H.T., Bradfield, J., Kim, C., Frackleton, E., Hou, C.P., Glessner, J.T., Chiavacci, R., Stanley, C., Monos, D., Grant, S.F.A., Polychronakos, C. and Hakonarson, H. (2009) 'From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes', *Plos Genetics*, Vol. 5.

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P.M., Baldassano, R.N., Hakonarson, H. and Consortium, I.I.G. (2013) 'Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease', *American Journal of Human Genetics*, Vol. 92, pp.1008–1012.

Welter, D., Macarthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. and Parkinson, H. (2014) 'The NHGRI GWAS Catalog, a curated resource of SNP-trait associations', *Nucleic Acids Research*, Vol. 42, pp.D1001–D1006.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) 'Rare-variant association testing for sequencing data with the sequence kernel association test', *The American Journal of Human Genetics*, Vol. 89, pp.82–93.

Yoon, D., Kim, Y.J. and Park, T. (2012) 'Phenotype prediction from genome-wide association studies: application to smoking behaviors', *Bmc Systems Biology*, Vol. 6.

Zou, H. and Hastie, T. (2005) 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B-Statistical Methodology*, Vol. 67, pp.301–320.