Genetics: Early Online, published on September 12, 2017 as 10.1534/genetics.117.300283

Improving disease prediction by incorporating family disease history in risk prediction models with large-scale genetic data

Jungsoo Gim[*], Wonji Kim[†], Soo Heon Kwak[‡], Hosik Choi[§], Changyi Park[**], Kyong Soo Park[††], Sunghoon Kwon[‡‡], Taesung Park[†,***], and Sungho Won[*,†,†††]

[*]Institute of Health and Environment, Seoul National University, Seoul 08826, Republic of Korea
Email: jgim80@snu.ac.kr

[†]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul 08826, Republic of Korea
Email: dnjswlzz@snu.ac.kr

[‡]Department of Internal Medicine, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
Email: shkwak@snu.ac.kr

[§]Department of Applied Information Statistics, Kyonggi University, Suwon, Republic of Korea
Email: choi.hosik@gmail.com

[**]Department of Statistics, University of Seoul, Seoul, Republic of Korea
Email: park463@uos.ac.kr

[††]Department of Internal Medicine, Seoul National University College of Medicine, Seoul 03080, Republic of Korea
Email: kspark@snu.ac.kr

[‡‡]Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea
Email: shkwon0522@gmail.com

[***]Department of Statistics, Seoul National University, Seoul 08826, Republic of Korea
Email: tspark@snu.ac.kr (corresponding author)

[†††]Graduate School of Public Health, Seoul National University, Seoul 08826, Republic of Korea
Email: won1@snu.ac.kr (corresponding author)

Abstract

Despite the many successes of genome-wide association studies (GWAS), the known susceptibility variants identified by GWAS have modest effect sizes, leading to notable scepticism about the effectiveness of building a risk prediction model from large-scale genetic data. However, in contrast to genetic variants, the family history of diseases has been largely accepted as an important risk factor in clinical diagnosis and risk prediction. Nevertheless, the complicated structures of the family history of diseases have limited their application in clinical practice. Here, we developed a new method that enables incorporation of the general family history of diseases with a liability threshold model, and propose a new analysis strategy for risk prediction with penalized regression analysis that incorporates both large numbers of genetic variants and clinical risk factors. Application of our model to type 2 diabetes (T2D) in the Korean population (1846 cases and 1846 controls) demonstrated that single nucleotide polymorphisms accounted for 32.5% of the variation explained by the predicted risk scores in the test data set, and incorporation of family history led to an additional 6.3% improvement in prediction. Our results illustrate that the family

medical history is valuable information on the variation of complex diseases and improves prediction performance.

Introduction

Despite the existence of promising examples of GWAS findings that will or may soon be translated into clinical utility (MANOLIO 2013), many studies have shown that genetic screening to predict the risk of complex diseases currently has little value in clinical practice (LYSSENKO AND LAAKSO 2013) and only shows modest predictive power even if all relevant loci (including rare variants) were discovered (CLAYTON 2009). For example, heritability estimates of type 2 diabetes (T2D) from twin and familial studies range from 40% to 80% (1988; KAPRIO *et al.* 1992), whereas the estimated heritability proportions explained by known susceptibility variants of T2D range from only 10% to 28%, indicating that most of the heritability remains unexplained (MCCARTHY 2010; SO *et al.* 2011a; SO *et al.* 2011b). In addition to this so-called 'missing-heritability' issue, GWAS-based common variants tend to only mildly predispose a carrier to a common disease (WEI *et al.* 2009), which generates some doubt about the overall value application of GWAS findings for risk assessment in clinical care (MANOLIO 2010).

The most popular approaches for disease risk prediction involve logistic regression analysis with genotype scores. With a training set, the regression coefficients of some significantly associated single nucleotide polymorphisms (SNPs) (MIYAKE *et al.* 2009) are calculated, and the sums of the weighted genotype scores with their regression coefficients are incorporated as a single covariate to the logistic regression for the test set (EVANS *et al.* 2009). However, the accuracy of these disease risk prediction models is generally much lower than what heritability estimates can provide.

To better convert heritability into prediction, several approaches have been proposed to include a large number of SNPs into the prediction model, including the use of penalized regression methods (WU *et al.* 2009; WON *et al.* 2015) and random-effects models (SPEED AND BALDING 2014). However, these attempts still have several limitations. The computational complexity linearly or quadratically increases with the number of SNPs depending on the algorithms, especially for the penalization approaches (WON *et al.* 2015). To reduce computational cost, it might be helpful to adopt a SNP-filtering strategy (filtering out less informative SNPs before building models). However, the performance of a prediction model based on this strategy largely depends on SNP-filtering methods.

An alternative is to incorporate family history. Family history reflects genetic susceptibility in addition to interactions between genetic, environmental, cultural, and behavioural factors (MACINNIS *et al.* 2011; DO *et al.* 2012). Therefore, it has been repeatedly suggested that incorporation of family medical history to a risk prediction model might implicitly cover the effects of uncovered genetic risk factors and shared gene-environment interactions (HARIRI *et al.* 2006; CHENG *et al.* 2015). Accordingly, family history has been often expected as an important risk factor in clinical assessment (HARIRI *et al.* 2006). Moreover, a recent theoretical work shows that including family history decreases the effective population size in prediction designs, thus resulting in a higher prediction accuracy (LEE *et al.* 2017). However, in spite of the known importance of family history, it is generally measured by an indicator

variable (showing the existence of known affected relatives) and this simple indicator has been incorporated into the prediction models. There is usually a great amount of heterogeneity among subjects with respect to familial relationships of relatives with known disease status, which has thus far limited the utility of this simplified binary variable for disease prediction.

In this article, we propose a new disease risk prediction model based on penalized regression with the following features: (i) a certain number of SNPs selected according to the absolute value of the best linear unbiased prediction (BLUP), (ii) penalized logistic regression analyses were performed using a number of SNPs leaving important predictive clinical variables un-penalized, and (iii) a new method is applied to incorporate the general family history of diseases. Application of our model to T2D patients in a Korean population showed that incorporation of family history could improve the amount of variation explained in the model. The model and approach proposed highlight the importance of family history of diseases for disease prediction, and is expected to become a useful tool to explain the variation of complex diseases.

Methods

In this section, we first introduce the process by which we evaluated a subject's conditional mean of disease risk using his/her family history and pre-screened SNPs based on the BLUP (Figure 1). With these variables, we then present how sparse modelling can be applied to build a risk prediction model for complex disease. Finally, we introduce two SNP chip datasets used in this study and propose a method of estimating the variance for each variable in the prediction model.

Evaluating the conditional mean of disease risk using family history

Suppose there are *n* subjects whose genotypes are known and each subject i has n_i $(i = 1, \dots, n)$ relatives whose genotypes are unknown, while disease status and relationship with subject i are available. We began our model by evaluating the conditional mean of disease risk using the standard liability threshold model (FALCONER 1967). We assume that disease status are determined by the unobserved liabilities (denoted as L), and if they are larger than a threshold *T*, which is determined by the disease prevalence, a subject will become affected. We further assume that these liabilities are normally distributed. Here, $Y_i = \left(Y_i, Y_{i_1}, \dots, Y_{i_{n_i}}\right)^t$, $L_i = \left(L_i, L_{i_1}, \dots, L_{i_{n_i}}\right)^t$ and $W_i = \left(W_i, W_{i_1}, \dots, W_{i_{n_i}}\right)^t$ respectively represent phenotypes, liabilities, and environment vectors of subject *i* and his/her family. For a given subject, we only use phenotypic information from their relatives, and we use subscript i_j to indicate relative j of subject i. We further denote by f_j and $\psi_{jj'}$ the inbreeding coefficient for relative j of subject i and the kinship coefficient between two relatives j and j' of subject i, respectively. It should be noted that $\psi_{jj'}$ is 0 if subjects j and j' are in different families. We then define the kinship coefficient matrix as Ψ_i, where $(\Psi_i)_{jj'}$ is $2\psi_{jj'}$ for $j \neq j'$, and is $1 + f_j$ otherwise. We denote a $k \times k$ dimensional identity matrix by I_k, and k dimensional column vector of which all elements are 0 by 0_k. With these notations, we assume that

$$L_i = W_i\alpha + P_i + E_i \quad (1)$$

$$P_i \sim MVN\left(0_{n_i+1}, \sigma_g^2 \Psi_i\right), E_i \sim MVN\left(0_{n_i+1}, \sigma_\epsilon^2 I_{n_i+1}\right) \quad (2)$$

where α indicates coefficient vector of fixed effects and σ_g^2 and σ_ϵ^2 indicate the variances of the random polygenic effect and random residual effect, respectively.

As a way to include family history, we introduced a variable, conditional mean (CM), which is defined as an individual's expected liability given case-control status of their relatives (KIM *et al.* 2017); CM reflects the likelihood of an individual to develop the disease (here, T2D) using only the knowledge of whether their relatives had the disease. Then, when comparing different regression methods, we included CM as we do other clinical covariates. To evaluate CM for each individual, it is necessary to integrate over all possible liabilities for the remaining individuals. Given an individual's case/control status, their liability takes a truncated normal distribution, so CM represent an expectation of a multivariate truncated normal distribution, which can be evaluated in R by using *pmvnorm*() function in *mvtnorm* package (see Supplementary Note for detailed methods). We implemented our method of evaluating familial risk in R and the source code and R package are available on Github (https://github.com/JungsooGim/familyRisk).

SNP pre-screening

To select an effective list of SNPs to test the model, we considered the BLUP of SNP effects using GCTA (YANG *et al.* 2011), which is a mixed linear model with the random effects of SNPs; i.e., $Y = W\beta + Gb + \epsilon$ with $b \sim MVN(0_n, \sigma_g^2 I_n)$ and $\epsilon \sim MVN(0_n, \sigma_\epsilon^2 I_n)$, thus leading to the mixed model $Y \sim MVN(W\beta + \sigma_g^2 GG' + \sigma_\epsilon^2 I_n)$. Here, G is a genotype matrix in training sets. The variance components σ_g^2 and σ_ϵ^2 are solved using restricted maximum likelihood (REML), which also provides an estimate of each individual genetic random effect, b. From this, the BLUP of SNP effects can be obtained via $\hat{b} = G'K^{-1}(Y - W\hat{\beta})/\hat{\sigma}_g^2$, and it can be simply obtained with GCTA (YANG *et al.* 2011). We ranked SNPs based on the absolute value of these estimated SNP effects. We also selected a list of SNP based on the p-value from the univariate logistic regression for each SNP with age, sex, body mass index (BMI), systolic blood pressure (SBP), and diastolic blood pressure (DBP) adjusted.

Penalized regression method

Let $X_i = (Z_i, W_i)$ and y_i be a covariate vector and a dichotomous phenotype for subject i, and affected and unaffected subjects are coded as 1 and 0, respectively. We further denote W_{il} and Z_{im} as coded genotypes of the lth SNP and the mth clinical covariate, respectively. The p-dimensional coefficient vector β consists of p_1 genetic variants and p_2 clinical variables. Under this model, β can be estimated by minimizing the penalized negative log-likelihood:

$$\frac{1}{n}\sum_{i=1}^{n}\{-y_i X_i'\beta + log(1 + exp(X_i'\beta))\} + \sum_{l=1}^{p_1} J_\lambda(|\beta_l|) \quad (3)$$

where J_λ is a penalty function and λ is a vector of a tuning parameter that can be determined by a search on an appropriate grid. Note that only genetic variants are penalized in Eq. 3.

For model analysis, Lasso (TIBSHIRANI 1996), Ridge (HOERL 1970), Elastic-Net (EN) (ZOU AND HASTIE 2005), SCAD (FAN AND LI 2001), and Truncated Ridge (TR) (CHATTERJEE AND LAHIRI 2011) can be performed depending on the choice of penalty function. Lasso, Ridge and EN were analysed under the default settings of *glmnet* (FRIEDMAN *et al.* 2010). For SCAD, whose penalty is defined as $\frac{\partial J_\lambda(t)}{\partial t} = \min\left\{\lambda, \frac{(a\lambda - t)_+}{a-1}\right\}$, we used $a = 50$ for our own optimization algorithm. For TR estimates, we first obtained ridge estimates with tuning parameter λ and then truncated them with a level a, so that the coefficients with absolute values smaller than a are set to zero. For the appropriate choice of a truncating level, 20 grid points (similar to EN) equally spaced in logarithmic scale from minimum to maximum ridge estimates were considered for a. All analyses were performed using R.

Building a disease risk model using the penalized regression method
In this section, we describe the brief steps for developing a disease risk model with the estimated CM score.

1. *Covariates*: Age, sex, body mass index (BMI), systolic blood pressure (SBP), and diastolic blood pressure (DBP) are considered as clinical covariates, and are included for all regressions.
2. *Summarising family history*: Calculate CM for all subjects with a familial history of disease.
3. *Data preparation for cross-validation*: Conduct 10-fold cross-validation. That is, the dataset is divided into 10 different sub-datasets, one of which is used as a test set and the other nine are used as training sets.
4. *SNP screening for the prediction mode*l: Using the training set in each cross-validation replicate, SNPs are pre-screened with the different criteria (p-value and BLUP) as described in SNP pre-screening in the previous section of Methods. For p-value criteria, SNPs with the top-*k* smallest p-values are selected. For BLUP criteria, SNPs with the top-*k* largest absolute BLUP values are selected. Here, we considered *k* = 100, 500, 1000, 5000, 10000, 20000.
5. *Model building*: Perform Lasso (TIBSHIRANI 1996), Ridge (HOERL 1970), EN (ZOU AND HASTIE 2005), SCAD (FAN AND LI 2001), and TR (CHATTERJEE AND LAHIRI 2011) for penalized regression. Tuning parameters for each penalized regression are selected with an additional 10-fold cross-validation using the training set. The training set is divided into 10 different sub-datasets, and for different choices of tuning parameters, the prediction model is obtained with the other nine sub-datasets. The area under the curve (AUC) is then calculated with the remaining sub-dataset, and tuning parameters that result in the largest AUC are finally chosen.
6. *Model validation*: The prediction models for penalized regressions are applied to the test set, and the AUCs are calculated.
7. *Performing cross-validation*: Repeat steps 4-6 for the different combinations of training and test sets.

Please note that ~285k genotyped SNPs were used in this work by assuming less chance of linkage disequilibrium (LD) among pre-screened SNPs in step 4. The assumption is likely to be falsified with a larger number of SNPs. In this case, an additional step of filtering SNPs within LD might be necessary.

Estimating variation in penalized logistic regression

To estimate the variation of each variable in the penalized regression model, we used the deviance calculated by comparing the predicted and true phenotypes in the test set. Specifically, we built the prediction model with a training set and the model was applied to predict the phenotypes of test samples. Then, the deviance was obtained by comparing the predicted phenotypes and the true phenotypes for those samples. If we denote the predicted and the true phenotypes by $\widehat{\mu_i}$ and y_i, respectively, the deviance is defined as

$$\Delta = \sum_i \left\{ y_i \, log \frac{y_i}{\widehat{\mu_i}} + (1 - y_i) \, log \frac{1-y_i}{1-\widehat{\mu_i}} \right\} \quad (4)$$

We used 10-fold cross validation and the deviances for all subjects were evaluated by summing all deviances in the test set. Based on Eq. 4, we defined the variation explained by the current model (Δ_F) using McFadden's R^2 (McFadden, 1974):

$$1 - \frac{\Delta_F}{\Delta_0} \times 100$$

where Δ_0 is the deviance of the null model. Then the variation unexplained by the full model can be obtained by 1- McFadden's. If we denote the reduced model whose ith element is excluded by Δ_i, and further defined the relative deviance explained by the ith variables as

$$1 - \frac{\Delta_F}{\Delta_0} \times 100 - \left(1 - \frac{\Delta_i}{\Delta_0} \times 100\right) = \frac{\Delta_i - \Delta_F}{\Delta_0} \times 100 \quad (5)$$

Eq. 5 represents the relative deviance explained by the ith variables among total variation.

Data description

To demonstrate the validity of our proposed model and to illustrate its application to disease risk prediction, we investigated type 2 diabetes (T2D) from two real datasets: KARE (Korea Association REsource) and SNUH (Seoul National University Hospital). Among the disease traits in KARE, T2D has the most well-investigated familial information and additional T2D cases and their well-organized familial history were available from SNUH. We merged the two datasets by adjusting for a platform difference (matching SNPs existing in both platforms and imputing missing genotypes using Shapeit). Overall, we analysed the data of 3692 subjects (1846 cases and 1846 controls) with a total of 267,063 SNPs.

As a part of the Korean NIH's project, the KARE cohort was recruited to construct an indicator of diseases with a genetic component in an attempt to predict disease outbreaks. Genotype information of 8,842 participants was received from Korea Center for Disease Control and Prevention. For these participants, 440,794 SNPs were genotyped with the Affymetrix Genome-Wide Human SNP array 5.0 and 267,064 SNPs remained in our analyses after the following quality controls: (1) p-values for Hardy-Weinberg equilibrium of less than 10^{-5}, (2) genotype call rates less than 95%, and (3) minor allele frequencies less than 0.05. We also eliminated subjects with gender inconsistencies, whose identity in state was more than 0.8, or whose call rates were less than 95%. Participants were asked whether they have affected relatives and if so, their ages and familial relatedness. The family histories of diseases, including

T2D, are also available for KARE data. Finally, we randomly selected controls to achieve the same number of cases and controls, and thus used 1846 T2D cases (1,167 from KARE and 679 from SNUH) and 1846 randomly selected controls.

For SNUH data, T2D patients were diagnosed as T2D using the World Health Organization criteria for Seoul National University Hospital, and 681 subjects with a positive family history of diabetes in first-degree relatives were preferentially included. The family history of their relatives was based on the recall of the proband. However, family members were encouraged to perform a 75 g oral glucose tolerance test, and subjects that were positive for glutamic acid decarboxylase autoantibodies test were excluded. In total, the disease statuses of 7,825 relatives of 681 subjects were available, and 2,875 of these relatives of the subjects had T2D. T2D patients originally diagnosed from Seoul National University Hospital were genotyped with the Affymetrix Genome-Side Human SNP array 5.0, and 480,589 SNPs were obtained. The same conditions for quality control with KARE were applied, and two subjects and 213,526 SNPs were excluded. In total, 679 T2D patients with 267,063 SNPs were used for the analysis.

Since these two datasets might be genetically distinct, the prediction is unduly driven by inclusion of the SNUH cases (who unlike the KARE individuals) do not have matched controls. Thus, we checked whether this was the case by drawing the MDS plot and the MAF scatter plot and found no difference between these two datasets (Figues S1-2). We downloaded the data sets from www.nih.go.kr/NIH followed by an approval process from Korean NIH (contact to biobank@korea.kr for further information)

Results

Characteristics of the variables

The methodology described earlier for estimating the CM for all subjects using their relatives in a pedigree was applied to the real datasets. Figure 2 shows the characteristics of six covariates included in the prediction model. As shown in Figure 2A, the mean values of CM are not much different between T2D cases and controls. However, more subjects with T2D had a significantly higher CM value compared to control subjects (mean values for cases and controls are 0.12 and 0.03, respectively with $p < E\text{-}10$ from two-sample t-test). Note that the individual with no family history has CM = 0. Similarly, all other clinical covariates are also significantly different between cases and controls ($p < 0.05$). The boxplots of other clinical covariates between cases and control are shown in Figure 2B–F. We also investigated the characteristics of both sets of SNPs selected according to the p-value and BLUP criteria, and found a similar pattern (Figure S3).

Comparison of the performance of the tested models

The purpose of this work was to investigate the performance of prediction models using family medical history and to construct the best T2D risk prediction model. For this purpose, we compared the performance of five different penalized regression methods by varying the number of SNPs with different measures of family history. To compare the performance of CM with other methods, we considered an alternative method, counting a weighted mean (WM) number of affected relatives in each pedigree, e.g., if individual 1 had 6 relatives of which 3 had T2D, a score of 3/6 is assigned. The key finding was that family history played a critical role in risk

prediction for all methods (Figure 3 and Table S1). Note the performance tendency across different methods of considering family history, the number of SNPs and the penalty can be more readily seen by bar plots (Figure 3), while the specific value can be found in the table (Table S1).

In the majority of cases, TR and Ridge revealed higher prediction performance compared to the other methods. Interestingly, similar behaviour was observed between Ridge and TR, and between Lasso and EN. We also investigated the models with the SNPs filtered by p-value criterion and observed a similar result (Tables S2-3). For a small number of SNPs, use of the p-value criterion showed better performance. However, the difference became negligible (or even reversed in some cases) as the number of SNPs increased. Among all comparisons, the best performance (AUC = 0.736) was observed when using Ridge and TR with CM and 5,000 SNPs selected by the BLUP criterion (Table S1). The general performance of the model with the WM variable was much lower than that of the model with CM and the performance is slightly higher than that of the model without family history in terms of AUC (Figure 3). Note that the standard deviation of the AUC was also evaluated and the range was 0.012-0.037. The best AUC value we obtained here is similar to that obtained previously (AEKPLAKORN *et al.* 2006; LYSSENKO *et al.* 2008).

To further investigate the effect of family history variable without SNPs, we built the logistic regression models without any SNPs. Based on the nested 10-fold cross-validation scheme, which was applied in the building steps of our model, we measured the performance of the logistic regression model without and with CM, or WM. Without CM, the AUC value was 0.672, but increased to 0.676 with WM included. When CM was included, AUC was 0.730. This value is similar to the highest AUC (0.736) obtained with 5,000 additional SNPs. Taken together, the model including CM increased the prediction performance the best in terms of AUC.
We next measured the time taken for the analysis for each method, and the results are shown in Table 2. We used 10 cores (CPUs) of our computing system (Intel Xeon CPU E5-2620 v2 @ 2.10Ghz). As can be seen, Ridge was the fastest method and truncated ridge followed. EN and SCAD were too slow to process the large genetic datasets even in our computing system. Note that unstable server usage Might affect the time for each analysis but the tendency described here was steady.

Variation explained by each variable
To estimate the variation explained by each variable, we investigated the model with 5,000 SNPs selected by the BLUP (please see the supplementary File S2 for BLUP of the SNPs). As described in the Methods section, we fit the reduced model to evaluate the residual deviance of each variable, calculated by comparing the predicted and true phenotypes in the data data set and the overall results are shown in Figure 4. Among the variation of the model including only an intercept, 42.0% of the total variation was explained by the full model which includes all covariates (age, sex, BMI, SBP, DBP, CM and 5k SNPs) (Figure S4). The largest portion (58.0%) of the variation remained unexplained, indicating that the variables in the model are not sufficient to explain the data. The second largest portion (32.5%) was derived from the SNPs. Even though the prediction performance was not significantly increased with these SNPs, they nevertheless explained about one-third of the total variation. Genetic heterogeneity of the disease can be one possible reason for this seemingly inconsistent result. If a substantial proportion of variance is explained by a certain variable, it is

usually expected to have a high predictive power. However, the better model fit does not always lead to the higher prediction performance in the existence of genetic heterogeneity of the disease.

In contrast, CM, which showed a dramatic increase in the prediction ability based on the AUC value, explained only 6.3% of the total variation. We also analysed the model without incorporating family history to compare the effect of CM to the proportion of variation. We excluded CM variable in the final model, and repeated the analyses to generate a pie chart. We found a ~9% decrease (larger than the 6.3% CM proportion) from the total amount of variation explained with CM to that without (Figure S5).

Discussion

Previous studies have documented the effectiveness of combining many SNPs using regularization methods or incorporating family history in improving the prediction performance of disease risk (MACINNIS *et al.* 2011; DO *et al.* 2012; WON *et al.* 2015). However, these studies have either been one-sided designs or were not simultaneously focused on both sides; i.e., combining more SNPs and also incorporating family history. In this study, we tested the extent to which combining SNPs and incorporating family history could improve risk prediction, and applied this approach to a dataset including a group of T2D patients and controls. We first developed a method to estimate the conditional mean of being affected by a disease for subjects in a pedigree. We then compared the prediction performance of six different regularization methods using SNPs selected by the p-value obtained from logistic regression and the BLUP value obtained from a mixed-effects model. We adopted a nested cross-validation scheme, which is time-consuming but known to be more reliable (VARMA AND SIMON 2006), to select the model showing the best prediction performance. Finally, we suggest a new method for estimating a variation explained by each variable in penalized regression models with a binary outcome (e.g., a case-control study).

In virtually all cases, the inclusion of family history (evaluated as CM) in the model greatly improved the prediction performance, while inclusion of SNPs showed only slight improvement. This finding indicates that proper incorporation of family history tends to produce a more effective genetic or environmental influence on the prediction results. Therefore, these benefits gained from incorporating CM might address the need for more rigorous investigations of gene-gene or gene-environmental interaction effects across a wide range of complex diseases. More importantly, a recent theoretical work has shown that using family information in training data would reduce effective population size, which leads to a better prediction model (LEE *et al.* 2017).

It has been thought that penalized regression models using all SNPs simultaneously may provide optimal performance even though it is infeasible in many cases because of the computational difficulty. Our results revealed that the additional power improvement is almost negligible if the number of SNPs is sufficiently large, compared to the scenario where all SNPs are used. Therefore, we can conclude that a well-developed feature selection method with sufficiently large number of SNPs preserves the optimal prediction accuracy and is beneficial because computational burden can be drastically reduced (FANG *et al.* 2008). In this work, we identified the

best-performed 5000 SNPs pre-screened by the BLUP. The top 5,000 SNPs that were pre-screened by our BLUP-based selection method showed the highest AUC value. Note that we also analysed the prediction model including all (~300k) SNPs using MultiBLUP (SPEED AND BALDING 2014) and the AUC was about 0.6 regardless of CM inclusion. MultiBLUP assumes binary phenotypes as continuous values. This requires ridge penalty to be applied to the linear regression, not logistic regression, resulting the likelihood quite different. Logistic regression is usually similar to the linear regression if the predicted probability is around 0.5; yet, in our penalized regression model, there are many covariates that can push the probability away from 0.5, Thus, this may result from the less predictive performance of MultiBLUP. Interestingly, the variation explained by these BLUP-based 5,000 SNPs (32.5%) was similar to the variance estimated by all SNPs (35%) reported to date (SPEED *et al.* 2012).

However, there are some limitations of the study that are worth noting. First, we did not consider other types of structural variants such as copy number variations, which might also affect the risk of T2D and the specific contribution is starting to be reported (DAJANI *et al.* 2015). Second, it would be preferable to include rarer risk alleles with large effects and gene-gene or gene-environment interactions into the prediction model. More of the genetic risk can likely be explained as more causal risk variants are identified. However, rare variant analyses or interaction analyses require more complicated statistical methods to effectively analyse the effects. Also, gene-gene and gene-environmental interactions are important, but not clearly considered in this work. Note that the model might capture some familial phenotypic correlation due to environmental factors, which is likely proportional to kinship. Thus, genetic correlation can be distorted by environmental interaction. Therefore, the ultimate goal of future work is to integrate advanced statistical methods with accumulating genetic data, environmental effects, and biological knowledge to improve the efficiency of detecting complex interactions. In addition, no effect of LD among SNPs was considered here. However, if a prediction model includes many SNPs in high LD, the locus effect gets divided between many SNPs in LD and might affect the prediction performance, especially with the BLUP-filtering criterion. We assumed no such LD effect because a small set of SNPs were selected from only 285k genotyped SNPs and they were all distant to each other (Figure S3). However, an inclusion of the LD-pruning step is desirable with a larger number of genotyped SNPs or imputed SNPs. Also, there are number of measures of predictive performance, but we only considered AUC in this work. The main goal of this report was to compare the performance of the prediction model with and without family history (CM) `using the AUC, and was not carefully stressed. However, it should be noted that for practical purposes, it is generally recommended that a prediction model provides both positive and negative results with the optimal threshold.

Funding
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 (H I15C2165).

Conflict of Interest: none declared.

References

1988 Diabetes mellitus in twins: a cooperative study in Japan. Committee on Diabetic Twins, Japan Diabetes Society. Diabetes Res Clin Pract 5**:** 271-280.

Aekplakorn, W., P. Bunnag, M. Woodward, P. Sritara, S. Cheepudomwit *et al.*, 2006 A risk score for predicting incident diabetes in the Thai population. Diabetes Care 29**:** 1872-1877.

Chatterjee, A., and S. N. Lahiri, 2011 Bootstrapping Lasso Estimators. Journal of the American Statistical Association 106**:** 608-625.

Cheng, H., L. Treglown, S. Montgomery and A. Furnham, 2015 Associations between Familial Factor, Trait Conscientiousness, Gender and the Occurrence of Type 2 Diabetes in Adulthood: Evidence from a British Cohort. Plos One 10.

Clayton, D. G., 2009 Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5**:** e1000540.

Dajani, R., J. Li, Z. Wei, J. T. Glessner, X. Chang *et al.*, 2015 CNV Analysis Associates AKNAD1 with Type-2 Diabetes in Jordan Subpopulations. Sci Rep 5**:** 13391.

Do, C. B., D. A. Hinds, U. Francke and N. Eriksson, 2012 Comparison of family history and SNPs for predicting risk of complex disease. PLoS Genet 8**:** e1002973.

Evans, D. M., P. M. Visscher and N. R. Wray, 2009 Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 18**:** 3525-3531.

Falconer, D. S., 1967 The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Ann Hum Genet 31**:** 1-20.

Fan, J. Q., and R. Z. Li, 2001 Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96**:** 1348-1360.

Fang, L. H., Y. Lv and G. H. Du, 2008 [Progress in study of pharmacological effect of Cortex Fraxini]. Zhongguo Zhong Yao Za Zhi 33**:** 2732-2736.

Friedman, J., T. Hastie and R. Tibshirani, 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33**:** 1-22.

Hariri, S., P. W. Yoon, N. Qureshi, R. Valdez, M. T. Scheuner *et al.*, 2006 Family history of type 2 diabetes: a population-based screening tool for prevention? Genet Med 8**:** 102-108.

Hoerl, A. E., 1970 Ridge Regression. Biometrics 26**:** 603-&.

Kaprio, J., J. Tuomilehto, M. Koskenvuo, K. Romanov, A. Reunanen *et al.*, 1992 Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia 35**:** 1060-1067.

Kim, W., D. Qiao, M. H. Cho, S. H. Kwak, K. S. Park *et al.*, 2017 Selecting cases and controls for DNA sequencing studies using family histories of disease. Stat Med 36**:** 2081-2099.

Lee, S. H., W. M. Weerasinghe, N. R. Wray, M. E. Goddard and J. H. van der Werf, 2017 Using information of relatives in genomic prediction to apply effective stratified medicine. Sci Rep 7**:** 42091.

Lyssenko, V., A. Jonsson, P. Almgren, N. Pulizzi, B. Isomaa *et al.*, 2008 Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359**:** 2220-2232.
Lyssenko, V., and M. Laakso, 2013 Genetic screening for the risk of type 2 diabetes: worthless or valuable? Diabetes Care 36 Suppl 2**:** S120-126.
Macinnis, R. J., A. C. Antoniou, R. A. Eeles, G. Severi, A. A. Al Olama *et al.*, 2011 A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. Genet Epidemiol 35**:** 549-556.
Manchia, M., J. Cullis, G. Turecki, G. A. Rouleau, R. Uher *et al.*, 2013 The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. PLoS One 8**:** e76295.
Manolio, T. A., 2010 Genomewide association studies and assessment of the risk of disease. N Engl J Med 363**:** 166-176.
Manolio, T. A., 2013 Bringing genome-wide association findings into clinical use. Nat Rev Genet 14**:** 549-558.
McCarthy, M. I., 2010 Genomics, type 2 diabetes, and obesity. N Engl J Med 363**:** 2339-2350.
Miyake, K., W. Yang, K. Hara, K. Yasuda, Y. Horikawa *et al.*, 2009 Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. J Hum Genet 54**:** 236-241.
So, H. C., A. H. Gui, S. S. Cherny and P. C. Sham, 2011a Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol 35**:** 310-317.
So, H. C., J. S. Kwan, S. S. Cherny and P. C. Sham, 2011b Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet 88**:** 548-565.
Speed, D., and D. J. Balding, 2014 MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 24**:** 1550-1557.
Speed, D., G. Hemani, M. R. Johnson and D. J. Balding, 2012 Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91**:** 1011-1021.
Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological 58**:** 267-288.
Varma, S., and R. Simon, 2006 Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7**:** 91.
Wei, Z., K. Wang, H. Q. Qu, H. Zhang, J. Bradfield *et al.*, 2009 From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 5**:** e1000678.
Won, S., H. Choi, S. Park, J. Lee, C. Park *et al.*, 2015 Evaluation of Penalized and Nonpenalized Methods for Disease Prediction with Large-Scale Genetic Data. Biomed Res Int 2015**:** 605891.
Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel and K. Lange, 2009 Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25**:** 714-721.
Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88**:** 76-82.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B-Statistical Methodology 67**:** 301-320.

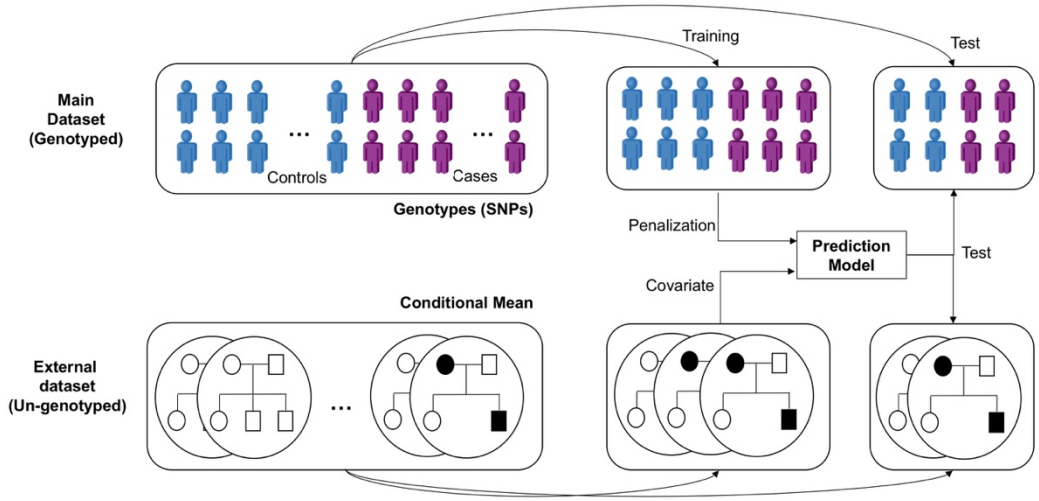

Figure 1 A schematic overview of the analysis. Individuals in the main dataset include genotyped SNPs, while the external dataset of those individuals includes relative's relationship and disease status. A 10-fold cross validation scheme was applied to build and test the performance of the prediction models

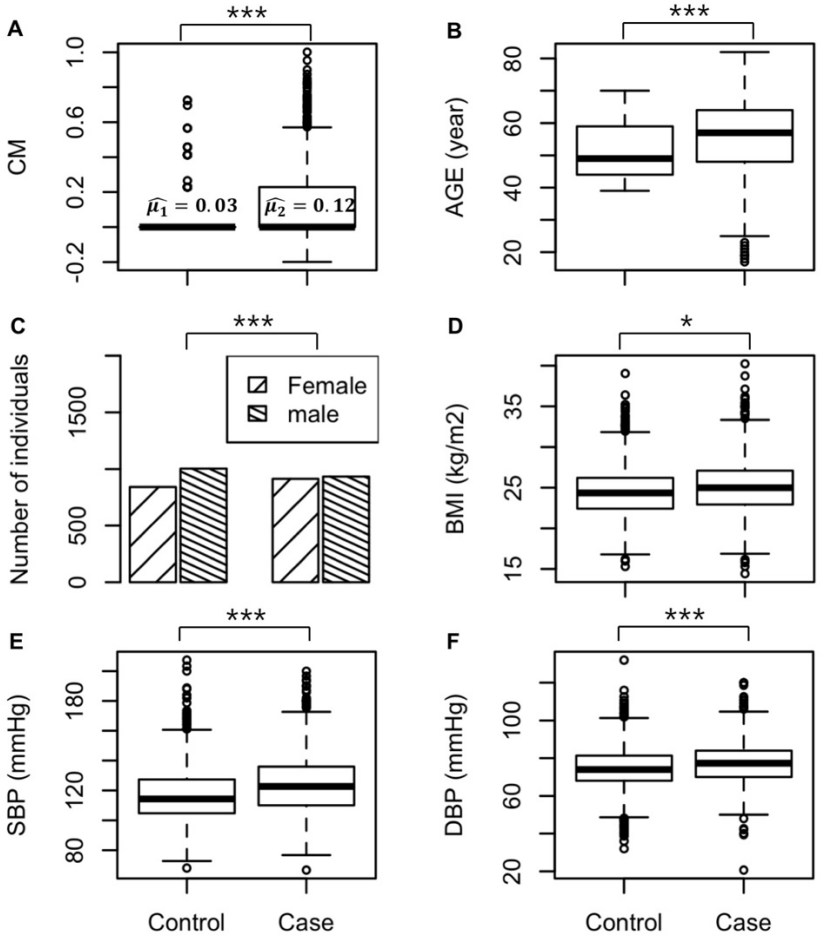

Figure 2 Clinical variables between cases and controls. Conditional mean (CM; A), age (B), sex (C), body mass index (BMI; D), systolic blood pressure (SBP; E), and diastolic blood pressure (DBP; F) are shown in boxplots. Two-sample t-test was performed to obtain p-values. For sex, a $\chi 2$-test was conducted.

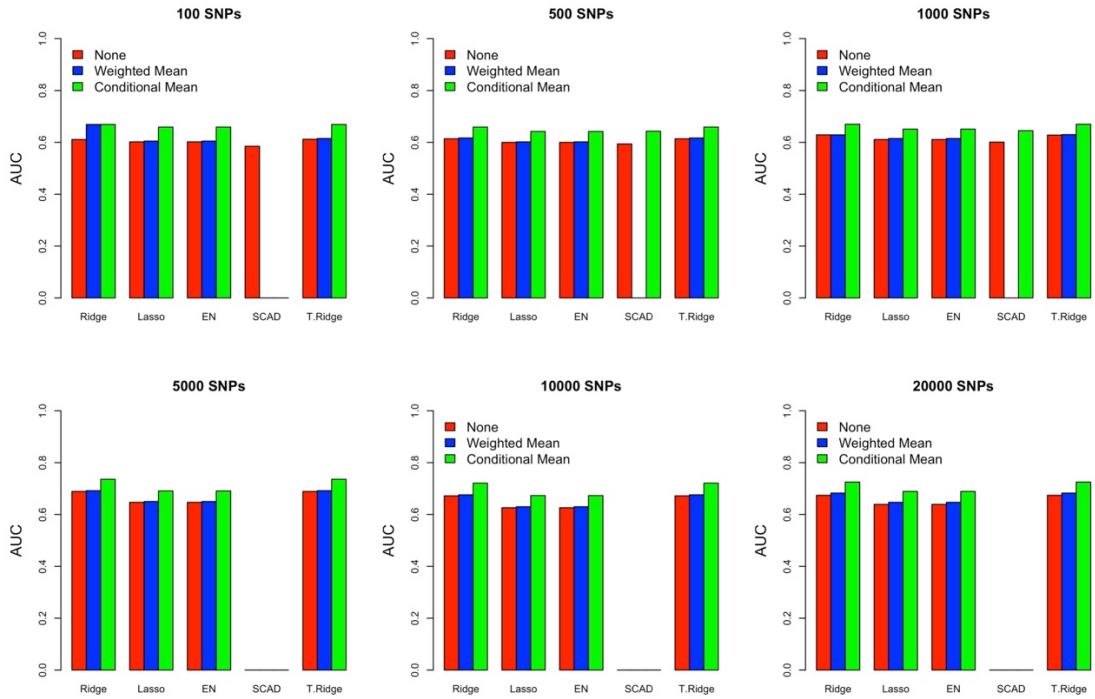

Figure 3 Model comparison with different family history measures. Prediction performance (AUC) is depicted without family history (red), with weighted mean (blue) and with conditional mean (green).

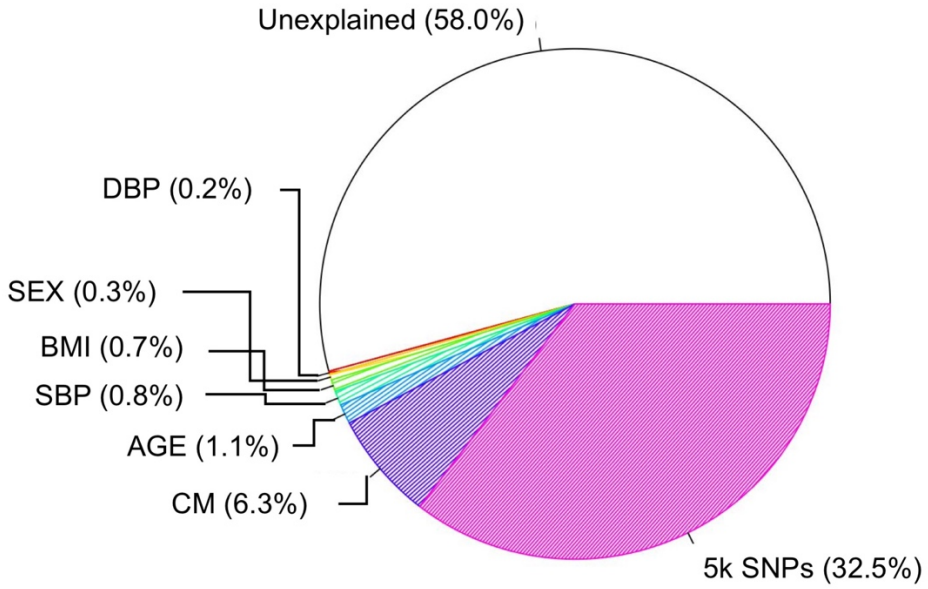

Figure 4 Proportion of variation explained by each variable in the final model. For six clinical variables (age, sex, BMI, SBP, DBP, CM), the individual proportions of variation are shown, whereas variation explained by the 5,000 SNPs is shown according to their summed proportion.

Table 1 Analysis Time

# of SNPs	RIDGE	LASSO	EN	SCAD	T.RIDGE
100	15.6 sec	13.2 sec	4.7 min	37 min	1.9 min
500	1.2 min	1.2 min	25.1 min	5.2 hour	6.0 min
1,000	2.6 min	2.2 min	43.5 min	12.2 hour	11.1 min
5,000	12.3 min	53.7 min	1.2 days	~ 3 days*	34.4 min
10,000	24.3 min	1.7 hour	2.3 days	~ 6 days*	1.7 hour
20,000	47.7 min	3.4 hour	3.61 days	~ 12 days*	3.3 hour

*Not measured but estimated