## TITLE:

Incorporating family history of disease into the prediction model with large-scale genetic data can dissolve the missing heritability of complex diseases

## AUTHORS:

Jungsoo Gim[*]
*Institute of Health and Environment*
*Seoul National University, South Korea*
*(jgim80@snu.ac.kr)*

Wonji Kim[*]
*Interdisciplinary Program of Bioinformatics*
*Seoul National University, South Korea*
(dnjswlzz@snu.ac.kr)

Soo Heon Kwak
*Department of Internal Medicine*
*Seoul National University College of Medicine, South Korea*
*(shkwak@snu.ac.kr)*

Kyong Soo Park
*Department of Internal Medicine*
*Seoul National University College of Medicine, South Korea*
*(kspark@snu.ac.kr)*

Sunghoon Kwon[¶]
*Department of Applied Statistics*
*Konkuk University, South Korea*

Sungho Won[¶]
*Graduate School of Public Health*
*Seoul National University, South Korea*
*(won1@snu.ac.kr)*

[*] Equally contributed authors
[¶]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 skepticism 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 to clinical use. Here, we developed a new method that enables the 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-scale genetic variants and clinical risk factors. Application of our model to type 2 diabetes (T2D) patients in the Korean population (1846 cases out of 3692 subjects) demonstrated that single nucleotide polymorphisms accounted for 28.6% of the variability of risk in T2D cases, and incorporation of family history led to an additional 5.9% improvement in prediction. Our results illustrate that the family history of diseases represents invaluable information for disease prediction and may bridge the missing heritability gap.

## INTRODUCTION

Despite the success in translating the significant results from genome-wide association studies (GWAS) to clinical utility(Manolio, 2013), many studies have shown that genetic screening for the prediction of complex diseases currently has little value in clinical practice(Lyssenko and Laakso, 2013). 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 proportions of the heritability explained by known susceptibility variants of T2D range from only 10% to 27.93%, indicating that most of the heritability is still unexplained(McCarthy, 2010; So, et al., 2011; So, et al., 2011). 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 of GWAS findings for risk assessment in clinical care(Manolio, 2010).

Alternatively, family history reflects genetic susceptibility in addition to interactions between genetic, environmental, cultural, and behavioral factors(Do, et al., 2012; Macinnis, et al., 2011). Therefore, it has been repeatedly suggested that incorporation of the family history of diseases to a risk prediction model might implicitly cover the effects of uncovered genetic risk factors and shared gene-environment interactions(Cheng, et al., 2015; Hariri, et al., 2006). According, family history has been often expected as an important risk factor in clinical assessment(Hariri, et al., 2006).

There have been many investigations for disease risk prediction based on large-scale genetic data and family history of diseases. 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 such disease risk prediction models is generally much lower than that expected from the heritability estimates. To overcome the controversy over the potential clinical value of GWAS findings, several approaches have been proposed to include a large number of SNPs into the prediction model, including the use of penalized regression methods(Won, et al., 2015; Wu, et al., 2009) and random-effects models(Speed and Balding, 2014). However, these attempts still have several limitations. For the penalized approaches, the computational intensity linearly or quadratically increases with the number of SNPs(Won, et al., 2015). Therefore, the accuracy of a prediction model based on penalized regression depends on the initial feature screening step, because a certain number of SNPs has to be chosen among the SNPs showing marginal effects, and those showing joint effects are ignored for feature selection. Speed et al. solved this problem with a random-effects model for linear regression in which disease status was considered a continuous response variable. However, in such a case, substantial bias can be observed if the probability of being affected is very small or large(Speed and Balding, 2014).

In this report, we propose a new disease risk prediction model based on penalized regression with the following features: (i) a certain number of SNPs are selected according to the best linear unbiased prediction (BLUP), (ii) penalized logistic regression analyses are performed using both SNPs and clinical variables, and (iii) a new method is applied to incorporate the general family history of diseases. However, in spite of the known importance of family history, 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 variable for disease prediction models. Application of our model to T2D patients in a Korean population showed that incorporation of family history could improve the amount of variability 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 bridge the gap derived from missing heritability.

## METHODS

### Evaluating the posterior mean of disease risk of a subject using family history

We assume that genotypes are not used to estimate the posterior mean of disease risk and that environmental effects are known. We began our model by evaluating the posterior mean of disease risk using the standard liability threshold model(Falconer, 1967). We assume that disease statuses 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, $\boldsymbol{Y}_i = \left(Y_{i_0}, Y_{i_1}, \dots, Y_{i_{n_i-1}}\right)^t$, $\boldsymbol{L}_i = \left(L_{i_0}, L_{i_1}, \dots, L_{i_{n_i-1}}\right)^t$, and $\boldsymbol{Z}_i = \left(Z_{i_0}, Z_{i_1}, \dots, Z_{i_{n_i-1}}\right)^t$ respectively represent phenotypes, liabilities, and environment vectors of subject *i* and his/her family. The number of relatives of subject $i$ is denoted as $n_i$. For a given subject, we only need information of the relatives of that subject, thus we use subscript $i_j$ to indicate each family member of subject $i$ ($j = 0$ indicates the subject $i$ itself). 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 $\boldsymbol{\Psi}_{\mathrm{i}}$, where $(\boldsymbol{\Psi}_{\mathrm{i}})_{jj'}$ is $2\psi_{jj'}$ for $j \neq j'$, and is $1 + f_j$ otherwise. With this notation, we assume that

$$\boldsymbol{L}_{i_j} = \boldsymbol{Z}_{i_j}\alpha + \boldsymbol{P}_{i_j} + \boldsymbol{E}_{i_j}, \boldsymbol{P}_{i_j} \sim MVN\left(\mathbf{0}_{n_i}, \sigma_g^2 \boldsymbol{\Psi}_{\mathrm{i}}\right), \boldsymbol{E}_{i_j} \sim MVN\left(\mathbf{0}_{n_i}, \sigma_\epsilon^2 \boldsymbol{I}_{n_i}\right) \qquad \text{(Eq.1)}$$

where $\boldsymbol{I}_{n_i}$ is the $n_i \times n_i$ dimensional identity matrix, and $\mathbf{0}_{n_i}$ and $\mathbf{1}_{n_i}$ are $n_i$-dimensional column vectors. Here, $\sigma_g^2$ and $\sigma_\epsilon^2$ indicate the variances of the polygenic effect and random effect, respectively.

Based on this liability threshold model, we can calculate the conditional expectation of $L_i$, PM, when the family histories of the disease are conditioned. We further define a random variable $\boldsymbol{A}_i$ of subject *i* by

$$\boldsymbol{A}_i = \left(A_{i_0}, \boldsymbol{A}_{i_j}\right)^t, A_{i_j} = \begin{cases} (T, \infty) & \text{if } Y_{i_j} = 1 \\ (-\infty, T) & \text{if } Y_{i_j} = 0 \end{cases} \text{ for } j = 0, \dots, n-1.$$

let $I_{A_{i_j}}\left(L_{i_j}\right) = 1$ if $L_{i_j} \in A_{i_j}$ and 0 otherwise, and $\mathrm{I}_{A_i}(\boldsymbol{L_i}) = \left(\mathrm{I}_{A_0}\left(L_{i_0}\right), \dots, \mathrm{I}_{A_{n-1}}(L_{n-1})\right)^t$, then PM becomes

$$\mathrm{E}\left(L_i \middle| \mathrm{I}\left(\boldsymbol{L}_{(-i)}\right) = \mathbf{1}_{\mathrm{n}-1}\right).$$

PM can be calculated with the moment generating function (*mgf*) of the truncated multivariate normal distribution to calculate the conditional distribution. The joint probability density function (*pdf*) can be defined as

$$f(\boldsymbol{L}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}\boldsymbol{L}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{L}\right), \qquad \text{(Eq.2)}$$

where $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{L})$. Based on the conditional pdf of $\boldsymbol{L}$ given $\mathrm{I}(\boldsymbol{L}) = \mathbf{1}$ and some algebra, we obtain

$$m(\boldsymbol{t}) = \frac{\exp\left(\frac{\boldsymbol{t}^t\boldsymbol{\Sigma}\boldsymbol{t}}{2}\right)}{Pr(\mathrm{I}_{\boldsymbol{A}}(\boldsymbol{L}) = \mathbf{1})(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \int_A \exp\left(-\frac{1}{2}\boldsymbol{L}^t\boldsymbol{\Sigma}\boldsymbol{L}\right) d\boldsymbol{L} \qquad \text{(Eq.3)}$$

If we let $(\boldsymbol{\Sigma})_{jk} = \sigma_{jk}$ and $F_k(\mathrm{x})$ be the marginal pdf of $L_k$, the PM for subject $i$ can be obtained by

$$\mu_i = \frac{\partial m(\boldsymbol{t})}{\partial t_i} = \sum_{k=1}^{n} \sigma_{ik}\, F_k^* \qquad \text{(Eq.4)}$$

where

$$F_k^* = \begin{cases} F_k(T) - F_k(\infty) & \text{if } y_k = 1 \\ F_k(-\infty) - F_k(T) & \text{otherwise} \end{cases}. \qquad \text{(Eq.5)}$$

The derivation of $F_k$ requires the marginal pdf of the truncated multivariate normal distribution, as follows. First, we partition $\boldsymbol{L}$ into two parts, $L_i$ and $\boldsymbol{L}_{(-i)}$, and then $\boldsymbol{L}$ can be rewritten as

$$\boldsymbol{L} = \begin{pmatrix} L_i \\ \boldsymbol{L}_{(-i)} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ \boldsymbol{0}_{n-1} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{22} & \Sigma_{22} \end{pmatrix}\right) \qquad \text{(Eq.6)}$$

If we denote the lower and upper truncated point of $\boldsymbol{L}$ as $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively, then the truncated normal distribution function when $\boldsymbol{a} < L < \boldsymbol{b}$ becomes

$$f_\alpha\left(L_i, \boldsymbol{L}_{(-i)} = \boldsymbol{x}\right) = \alpha^{-1} f\left(\boldsymbol{L}_{(-i)} = \boldsymbol{x}\right) f\left(L_i | \boldsymbol{L}_{(-i)} = \boldsymbol{x}\right). \qquad \text{(Eq.7)}$$

Using the marginal pdf of $\boldsymbol{L}_{(-i)}$ at $\boldsymbol{L}_{(-i)} = \boldsymbol{x}$ and the fact that the conditional distribution of the normal distribution is normally distributed, one can easily show that $L_i | \boldsymbol{L}_{(-i)} = \boldsymbol{x}$ follows the normal distribution with $E\left( L_i | \boldsymbol{L}_{(-i)} = \boldsymbol{x}\right) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{x}$ and $Var\left( L_i | \boldsymbol{L}_{(-i)} = \boldsymbol{x}\right) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Accordingly, the multivariate marginal pdf of $\boldsymbol{L}_{(-i)}$ becomes

$$F_{\boldsymbol{L}_{(-i)}}(x) = \int_{a_i}^{b_i} \alpha^{-1} f\left(\boldsymbol{L}_{(-i)} = \boldsymbol{x}\right) f\left(L_i | \boldsymbol{L}_{(-i)} = \boldsymbol{x}\right) dL_i \qquad \text{(Eq.8)}$$

The integral can be readily computed by using conventional statistical software. For this purpose, we used the *pmvnorm()* function in the R package *mvtnorm*(Wilhelm and Manjunath, 2010).

### Prescreening based on the BLUP

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., $\boldsymbol{Y} = \boldsymbol{W\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}$ with $var(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{A}\sigma_g^2 + \boldsymbol{I}\sigma_\epsilon^2$, where $\boldsymbol{y}$ and $\boldsymbol{\beta}$ are a vector of phenotypes and the fixed effect of subjects with genotypes, respectively, and $\boldsymbol{g}$ and $\boldsymbol{\epsilon}$ are vectors of the total genetic effects of the subjects with $g \sim N(0, \boldsymbol{A}\sigma_g^2)$ and residual effects with $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I}\sigma_\epsilon^2)$. **A** is the genetic relationship matrix (GRM) among subjects. Here coefficient vector can be estimated by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{W}'\boldsymbol{V}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{V}^{-1}\boldsymbol{Y}$. By estimating the GRM from all the SNPs and considering the following equivalent model, the BLUP of $\boldsymbol{g}$ can be provided by the restricted maximum likelihood (REML) approach.

Consider a mathematically equivalent model, $\boldsymbol{Y} = \boldsymbol{W\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$ with $var(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{ZZ^t}\sigma_g^2 + \boldsymbol{I}\sigma_\epsilon^2$, where $\boldsymbol{b}$ is a vector of random effects with $\boldsymbol{b} \sim N(0, \boldsymbol{I}\sigma_u^2)$ and $\boldsymbol{Z}$ is a standardized genotype matrix. The GRM **A** can be defined by $\boldsymbol{ZZ^t}/p_1$, where $p_1$ is the number of SNPs. Since these two equations are mathematically equivalent, the BLUP of $\boldsymbol{g}$ can be transformed to the BLUP of $\boldsymbol{b}$ by using $\widehat{\boldsymbol{b}} = \boldsymbol{Z}'\boldsymbol{A}^{-1}\widehat{\boldsymbol{g}}/p_1$ or can be calculated directly using the equation $\widehat{\boldsymbol{b}} = \boldsymbol{Z}'V^{-1}\left(\boldsymbol{Y} - \boldsymbol{W}\widehat{\boldsymbol{\beta}}\ \right)/\sqrt{p_1}$. Thus, the estimate of $b_l$ corresponds to the coefficient $Z_{il}$, which is the $l$th SNP of the $i$th subject element of $\boldsymbol{Z}$. Note that $Z_{il} = (G_{il} - 2d_l)/\sqrt{2d_l(1-d_l)}$, where $G_{il}$ and $d_l$ are the numbers of copies of the reference allele and the frequency of the reference allele of the $l$th SNP, respectively. When divided by $\sqrt{2d_l(1-d_l)}$, $\widehat{b_l}$ can be rescaled for the original genotype. We evaluated the BLUP for all the SNPs using the GCTA tool.

### Penalized regression method

Let $\boldsymbol{X}_i = (\boldsymbol{Z}_i, \boldsymbol{W}_i)$ 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 $Z_{il}$

and $W_{im}$ as coded genotypes of the $l$th SNP and the $m$th clinical covariate, respectively. The $p$-dimensional coefficient vector $\boldsymbol{\xi} = \left(\xi_1, \dots, \xi_{p_1}, \xi_{p_{1+1}}, \dots, \xi_{p_1+p_2=p}\right)^t$ consists of $p_1$ genetic variants and $p_2$ clinical variables. Under this model, $\boldsymbol{\xi}$ can be estimated by minimizing the penalized negative log-likelihood:

$$\frac{1}{n}\sum_{i=1}^{n}\{-y_i\boldsymbol{X}_i'\boldsymbol{\xi} + \log(1 + \exp(\boldsymbol{X}_i'\boldsymbol{\xi}))\} + \sum_{l=1}^{p_1} J_\lambda(|\xi_l|) \qquad \text{(Eq.9)}$$

where $J_\lambda$ is a penalty function and $\lambda$ 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. 9.

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. The penalty of Lasso is $J_\lambda(t) = \lambda t$, which is often utilized because Lasso can conduct both shrinkage and variable selection. Even though Lasso has an overfitting problem, its performance is quite stable, especially when the sample size is small. Ridge uses $J_\lambda(t) = \lambda t^2$ as its penalty. Similar to Lasso, it exerts the shrinkage effect by choosing $\lambda$ but no other variables. Ridge can be applied even when $p$ is much larger than $n$. EN, which is a convex combination of Lasso and Ridge, has a penalty of $J_\lambda(t) = \lambda(\alpha t + (1-\alpha)t^2)$, and we considered 20 equally spaced grid points from 0 to 1 for $\alpha$. EN allows for balanced estimates, producing a slightly more complex model than Lasso but a far simpler model than ridge. The penalty of SCAD is $\frac{\partial J_\lambda(t)}{\partial t} = \min\left\{\lambda, \frac{(a\lambda - t)_+}{a-1}\right\}$, and we used $a = 50$ for our own optimization algorithm. SCAD is known to have the oracle property, i.e., the set of selected variables is asymptotically equal to the set of true causal variables. In spite of its theoretical optimality, SCAD estimates can be poor unless the sample size is large and the effects of the signal variables are strong(Won, et al., 2015). For TR estimates, we first obtained ridge estimates with tuning parameter $\lambda$ 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 equally spaced in logarithmic scale from minimum to maximum ridge estimates were considered for $a$. All analyses were performed with the *glmnet*(Friedman, et al., 2010) R package.

### 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 PM score.

1. 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. Calculate PM for all subjects with family histories of diseases.
3. 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. Using the training set, *k* SNPs are selected according to p-values showing marginal effects from logistic regression and from the proposed BLUP method. Here, we considered *k* = 100, 500, 1000, 5000, 10000, 20000.
5. 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, as well as the mixed-effects model (MultiBLUP(Speed and Balding, 2014)).

   Tuning parameters for each penalized regression are chosen 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. The prediction models for penalized regressions and multiBLUP are applied to the test set, and the AUCs are calculated.
7. Repeat steps 3–7 for the different combinations of training and test sets.

### Data Description

To demonstrate the validity of our proposed model and to illustrate its application to disease risk prediction, we investigated two real datasets: KARE and SNUH. Since the SNUH dataset includes cases only, we merged the two datasets by adjusting for a platform difference (matching SNPs existing in both platforms and imputing NAs using Shapeit). Overall, we analyzed the data of 3692 subjects (1846 cases and 1846 controls) with a total of 267,063 SNPs.

The KARE cohort was recruited to construct an indicator of diseases with a genetic component in an attempt to predict disease outbreaks. There were initially 8,842 participants who were genotyped for 352,228 SNPs with the Affymetrix Genome-Wide Human SNP array 6.0. In our study, SNPs with the following characteristics were discarded in further analysis: (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 the KARE data. Finally, 1,167 T2D cases and 1846 randomly selected controls with 267,063 SNPs were used for the analysis.

For the 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.

### Estimating variability in penalized logistic regression

To estimate the variability of each variable in the penalized regression model, we used residual deviance from the penalized log-likelihood. The residual deviance is defined as

$$\Delta_{\mathrm{res}} = -2\left(l_{penal}(\beta)\right) \qquad \text{(Eq.10)}$$

where $l_{penal}(\beta) = Y^t \log(P) + (1-Y)^t \log(1-P) - \frac{1}{2}\lambda\sum_{i=1}^{p}\beta_i^2$ and $\mathrm{P} = \frac{\mathrm{e}^{\mathrm{x}^{\mathrm{t}}\beta}}{1+\mathrm{e}^{\mathrm{x}^{\mathrm{t}}\beta}}$. Using Eq. 10, we defined the variability explained by the $i$th reduced model as

$$\frac{\left|\Delta_{\mathrm{res,i}} - \Delta_{res,0}\right|}{\Delta_{res,0}} \times 100 \qquad \text{(Eq.11)}$$

where $\Delta_{res,0}$ denotes the residual deviance of the null model.

## RESULTS

### Characteristics of the variables

The established methodology for estimating the PM for all subjects in a pedigree was conducted and the method was applied to the real datasets described above. As shown in Fig. 1A, the mean values of PM did not differ between T2D cases and controls. However, more subjects with T2D had a relatively high PM value (larger than 0.5) compared to control subjects. The boxplots of other clinical covariates between cases and control are shown in Fig. 1B–F.

To find the most effective set of SNPs, we selected SNPs based on the p-value obtained from the logistic regression and the BLUP obtained by the mixed-effects model. Since the selected set of SNPs should be applied in penalized regression, we expected that the selection procedure would be more effective if the set of SNPs was uniformly distributed across the genome. Toward this end, we discretized the whole genome with a window size of 5 M base pairs and counted the frequency of SNPs in each window. With a varying number of SNPs (0.1–20 k), both sets of SNPs selected according to the p-value and BLUP criteria exhibited similar patterns (result not shown).

### Comparison of the performance of the tested models

The main purpose of this work was to construct a T2D risk prediction model. To find the best model, we sought to compare the performances of six methods using different criteria for selecting SNPs and by varying the number of SNPs. We repeated our analysis with family history (Table I) and without family history (Table II). The most striking finding was that family history (PM) played a very important role in risk prediction for all methods, except when using MultiBLUP. By comparing Table I and II, it is obvious that a significant improvement was obtained with the prediction model using PM variables.

In the majority of cases, TR and Ridge revealed higher prediction performance compared to the other methods. Interestingly, similar behavior was observed between Ridge and TR, and between Lasso and EN. For a small number of SNPs, use of the p-value criteria showed better performance. However, the difference became negligible (or even reversed in some cases) as the number of SNPs increased.

The best performance (AUC = 0.736) was observed when using Ridge and TR with PM and 5,000 SNPs selected by the BLUP criteria (TR showed a slightly higher AUC in the order of 10 -5). The AUC value we obtained here is similar to the results obtained in previous studies.(Aekplakorn, et al., 2006; Lyssenko, et al., 2008) To investigate the effect of each variable, we built the logistic regression 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 PM. Without PM, the AUC value was 0.672, but increased to 0.730 with PM included (Table III). This value is similar to the highest AUC (0.736) obtained with 5,000 additional SNPs.

We next measured the complexity of each method with respect to the time required, and the results are shown in Table IV. In general, the analysis time increased if the number of SNPs increased, except with MultiBLUP, which requires several manual steps to perform a prediction analysis. Therefore, it was difficult to measure the exact time for the whole analysis. However, MultiBLUP was not substantially affected by the SNP number increment.

### Variability explained by each variable

To estimate the variability explained by each variable, we investigated the model with 5,000 SNPs selected by the BLUP. As described in the Methods section, we fit the reduced model to evaluate the residual deviance of each variable, and the overall results are shown in Figure 2. The largest portion (58.9%) of the variation remained unexplained, indicating that the variables in the model are not sufficient to explain the data. The second largest portion (28.6%) was derived from the SNPs. Even though the prediction performance was not significantly increased with these SNPs, they nevertheless explained about 30% of the total variability. In contrast, PM, which showed a dramatic increase in the prediction ability based on the AUC value, explained only 5.9% of the total variability.

## DISCUSSION & CONCLUSIONS

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(Do, et al., 2012; Macinnis, et al., 2011; 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 posterior 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 deriving heritability estimates for a binary outcome (e.g., a case-control study). Unlike previous methods such as GCTA, in which the heritability is estimated by assuming a binary response as a continuous variable, our suggested method uses the deviance of the models, and thus no constraint is assumed for a binary response variable.

In virtually all cases, the inclusion of family history (evaluated as the PM) 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 PM might address the need for more rigorous investigations of gene-gene or gene-environmental interaction effects across a wide range of complex diseases. We also found that including more SNPs in the analysis does not guarantee better performance for T2D prediction. The top 5,000 SNPs that were pre-screened by our BLUP-based selection method showed the highest AUC value. Interestingly, the variability (34.5%) explained by these BLUP-based 5,000 SNPs (28.9%) and PM (5.9%) was is 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 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 analyze the effects. Therefore, the ultimate goal of future work is to integrate advanced statistical methods with accumulating genetic data and biological knowledge, which will further improve the power to detect complex interactions efficiently.

## Acknowledgements

The authors greatly acknowledge Doug Speed and Christoph Lange for helpful discussions.

## Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2010437), by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028559) and by the Industrial Core Technology Development Program (10040176, Development of Various Bioinformatics Software using Next Generation Bio-data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (received by S.Won). This study was also supported by a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (grant no. HI14C0060) (received by S.-H. Kwak).

## FIGURE LEGENDS

**Figure 1. Clinical variables between cases and controls.** Characteristics of the PM (A), age (B), sex (C), BMI (D), SBP (E), and DBP (F) are shown in boxplots. Here, disease status 1 and 0 indicate T2D cases and controls, respectively.

**Figure 2. Proportion of variability explained by each variable in the final model.** For six clinical variables (age, sex, BMI, SBP, DBP, PM), the individual proportions of the variability are shown, whereas the variability explained by the 5,000 SNPs is shown according to their summed proportion.

## REFERENCES

Diabetes mellitus in twins: a cooperative study in Japan. Committee on Diabetic Twins, Japan Diabetes Society. *Diabetes research and clinical practice* 1988;5(4):271-280.

Aekplakorn, W., *et al.* A risk score for predicting incident diabetes in the Thai population. *Diabetes care* 2006;29(8):1872-1877.

Chatterjee, A. and Lahiri, S.N. Bootstrapping Lasso Estimators. *J Am Stat Assoc* 2011;106(494):608-625.

Cheng, H., *et al.* Associations between Familial Factor, Trait Conscientiousness, Gender and the Occurrence of Type 2 Diabetes in Adulthood: Evidence from a British Cohort. *Plos One* 2015;10(5).

Dajani, R., *et al.* CNV Analysis Associates AKNAD1 with Type-2 Diabetes in Jordan Subpopulations. *Sci Rep* 2015;5:13391.

Do, C.B., *et al.* Comparison of family history and SNPs for predicting risk of complex disease. *PLoS Genet* 2012;8(10):e1002973.

Evans, D.M., Visscher, P.M. and Wray, N.R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human molecular genetics* 2009;18(18):3525-3531.

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

Fan, J.Q. and Li, R.Z. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001;96(456):1348-1360.

Friedman, J., Hastie, T. and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010;33(1):1-22.

Hariri, S., *et al.* Family history of type 2 diabetes: a population-based screening tool for prevention? *Genet Med* 2006;8(2):102-108.

Hoerl, A.E. Ridge Regression. *Biometrics* 1970;26(3):603-&.

Kaprio, J., *et al.* Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* 1992;35(11):1060-1067.

Lyssenko, V., *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* 2008;359(21):2220-2232.

Lyssenko, V. and Laakso, M. Genetic screening for the risk of type 2 diabetes: worthless or valuable? *Diabetes care* 2013;36 Suppl 2:S120-126.

Macinnis, R.J., *et al.* A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. *Genet Epidemiol* 2011;35(6):549-556.

Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;363(2):166-176.

Manolio, T.A. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* 2013;14(8):549-558.

McCarthy, M.I. Genomics, type 2 diabetes, and obesity. *N Engl J Med* 2010;363(24):2339-2350.

Miyake, K*., et al.* Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. *Journal of human genetics* 2009;54(4):236-241.

So, H.C*., et al.* Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet Epidemiol* 2011;35(5):310-317.

So, H.C*., et al.* Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am J Hum Genet* 2011;88(5):548-565.

Speed, D. and Balding, D.J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome research* 2014;24(9):1550-1557.

Speed, D*., et al.* Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* 2012;91(6):1011-1021.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 1996;58(1):267-288.

Varma, S. and Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7:91.

Wei, Z*., et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* 2009;5(10):e1000678.

Wilhelm, S. and Manjunath, B.G. tmvtnorm: A Package for the Truncated Multivariate Normal Distribution. *R J* 2010;2(1):25-29.

Won, S*., et al.* Evaluation of Penalized and Nonpenalized Methods for Disease Prediction with Large-Scale Genetic Data. *Biomed Res Int* 2015;2015:605891.

Wu, T.T*., et al.* Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 2009;25(6):714-721.

Yang, J*., et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;88(1):76-82.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005;67:301-320.

## FIGURES & TABLES

**Table I.** AUC with clinical variables and SNPs

| CRITERIA | # of SNPs | RIDGE | LASSO | EN | SCAD | T.RIDGE | MultiBLUP |
|---|---|---|---|---|---|---|---|
| P-value | 100 | 0.642 | 0.637 | 0.637 | 0.616 | 0.641 | 0.532 |
| | 500 | 0.640 | 0.626 | 0.626 | 0.608 | 0.640 | 0.542 |
| | 1,000 | 0.640 | 0.624 | 0.624 | 0.608 | 0.640 | 0.544 |
| | 5,000 | 0.660 | 0.635 | 0.635 | - | 0.660 | 0.546 |
| | 10,000 | 0.668 | 0.640 | 0.640 | - | 0.668 | 0.560 |
| | 20,000 | 0.674 | 0.640 | 0.640 | - | 0.674 | 0.582 |
| BLUP | 100 | 0.611 | 0.602 | 0.602 | 0.585 | 0.612 | 0.500 |
| | 500 | 0.614 | 0.600 | 0.600 | 0.594 | 0.614 | 0.513 |
| | 1,000 | 0.626 | 0.611 | 0.611 | 0.601 | 0.626 | 0.537 |
| | 5,000 | **0.689** | 0.647 | 0.647 | - | **0.689** | 0.581 |
| | 10,000 | 0.672 | 0.626 | 0.626 | - | 0.672 | 0.550 |
| | 20,000 | 0.674 | 0.639 | 0.639 | - | 0.674 | 0.571 |

**Table II.** AUC with clinical variables, SNPs and PM

| CRITERA | # of SNPs | RIDGE | LASSO | EN | SCAD | T.RIDGE | MultiBLUP |
|---|---|---|---|---|---|---|---|
| P-value | 100 | 0.693 | 0.687 | 0.688 | 0.676 | 0.693 | 0.534 |
|  | 500 | 0.687 | 0.672 | 0.672 | 0.665 | 0.687 | 0.544 |
|  | 1,000 | 0.685 | 0.669 | 0.669 | 0.664 | 0.685 | 0.536 |
|  | 5,000 | 0.709 | 0.687 | 0.687 | - | 0.709 | 0.541 |
|  | 10,000 | 0.717 | 0.690 | 0.690 | - | 0.717 | 0.554 |
|  | 20,000 | 0.721 | 0.689 | 0.689 | - | 0.721 | 0.561 |
| BLUP | 100 | 0.669 | 0.659 | 0.659 | 0.643 | 0.669 | 0.500 |
|  | 500 | 0.659 | 0.642 | 0.642 | 0.639 | 0.659 | 0.505 |
|  | 1,000 | 0.670 | 0.651 | 0.651 | 0.645 | 0.670 | 0.516 |
|  | 5,000 | **0.736** | 0.691 | 0.691 | - | **0.736** | 0.575 |
|  | 10,000 | 0.721 | 0.673 | 0.673 | - | 0.721 | 0.544 |
|  | 20,000 | 0.725 | 0.689 | 0.689 | - | 0.725 | 0.562 |

**Table III.** AUC without SNPs

| Variables included | Logistic regression |
|---|---|
| Age, sex, SBP, DBP, BMI | 0.672 |
| Age, sex, SBP, DBP, BMI, PM | 0.730 |

**Table IV.** Analysis Time

| # of SNPs | RIDGE | LASSO | EN | SCAD | T.RIDGE | MultiBLUP |
|---|---|---|---|---|---|---|
| 100 | 15.6 sec | 13.2 sec | 4.7 min | 37 min | 1.9 min | < 20 min |
| 500 | 1.2 min | 1.2 min | 25.1 min | 5.2 hour | 6.0 min | < 20 min |
| 1,000 | 2.6 min | 2.2 min | 43.5 min | 12.2 hour | 11.1 min | < 20 min |
| 5,000 | 12.3 min | 53.7 min | 35.6 min | ~ 3 days* | 34.4 min | < 20 min |
| 10,000 | 24.3 min | 1.7 hour | 1.1 hour | ~ 6 days* | 1.7 hour | < 20 min |
| 20,000 | 47.7 min | 3.4 hour | 3.4 hour | ~ 12 days* | 3.3 hour | < 20 min |

*Not measured but estimated

## Figure 1

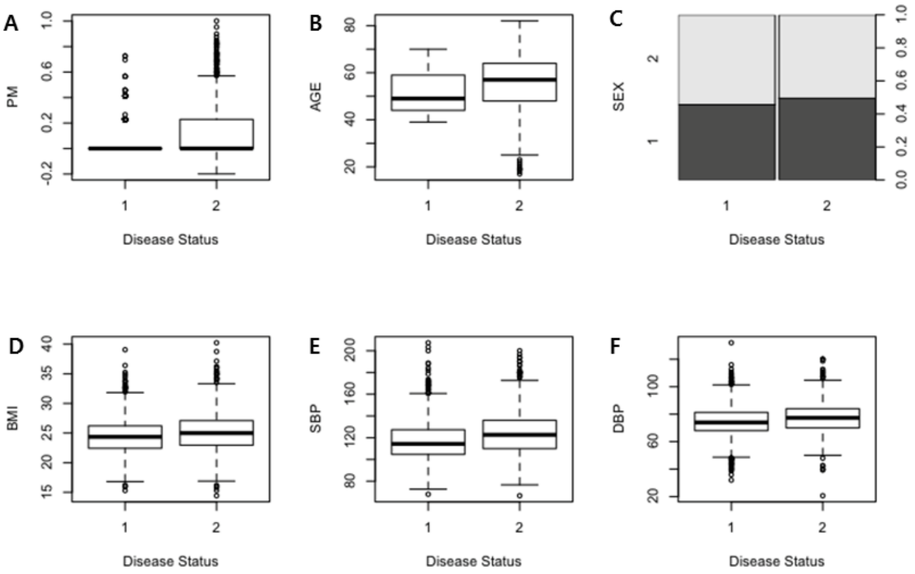
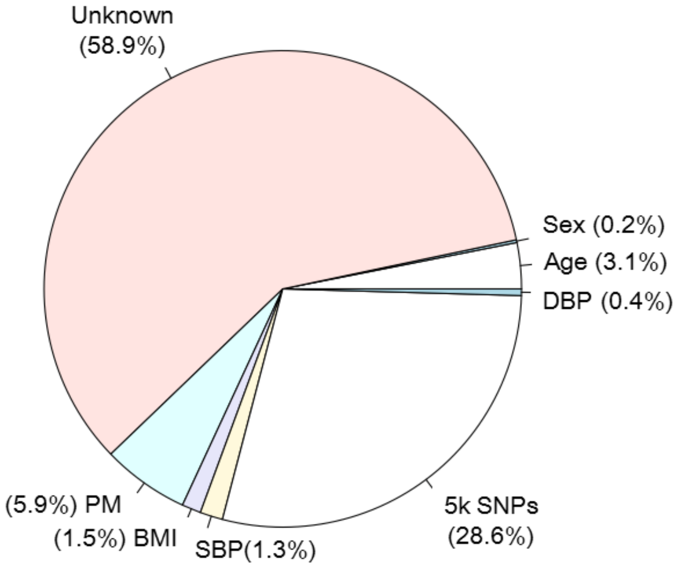

A
PM
1.0
0.6
0.2
-0.2
1
2
Disease Status
B
AGE
80
60
40
20
1
2
Disease Status
C
SEX
2
1
1.0
0.8
0.6
0.4
0.2
0.0
1
2
Disease Status
D
BMI
40
35
30
25
20
15
1
2
Disease Status
E
SBP
200
160
120
80
1
2
Disease Status
F
DBP
100
60
20
1
2
Disease Status

**Figure 2**


Unknown (58.9%)
Sex (0.2%)
Age (3.1%)
DBP (0.4%)
5k SNPs (28.6%)
SBP(1.3%)
(1.5%) BMI
(5.9%) PM