Supplementary note

A brief review of the liability threshold model

Consider n independent subjects with genotyped SNPs and their relatives whose relation with the subject and disease status and are only known. Phenotypes of individual i and his relative j are denoted by Y_i and Y_{i_j}, respectively and further define a vector $\mathbf{Y_i} = \left(Y_i, \boldsymbol{Y_{i_j}}\right)^{\boldsymbol{t}}$, where $\boldsymbol{Y_{i_j}} = \left(Y_{i_1}, \ldots, Y_{i_{n_i}}\right)^{\boldsymbol{t}}$. Similarly, the liability vector, the genotype vector and any environmental effects are denoted as $\mathbf{L_i} = \left(L_i, \boldsymbol{L_{i_j}}\right)^{\boldsymbol{t}}$, $\mathbf{G_i} = \left(G_i, \boldsymbol{G_{i_j}}\right)^{\boldsymbol{t}}$, and $\mathbf{Z_i} = \left(Z_i, \boldsymbol{Z_{i_j}}\right)^{\boldsymbol{t}}$ in order. Polygenic effects explain phenotypic similarity between family members, and correlations among family members are assumed to be constructed by kinship coefficients. We denote $\pi_{jj'}$ as the kinship coefficient between two relatives j and j' of subject i and d_{i_j} as the inbreeding coefficient for relative j of subject i. The inbreeding coefficient d_{ij} is a parameter quantifying the departure from HWE and ranges from 0 to 1. We denote the kinship coefficient matrix of subject i's relatives by

$$\boldsymbol{\Psi}_i^{\text{rel}} = \begin{pmatrix} 1 + d_{i1} & \cdots & 2\pi_{1n_i} \\ \vdots & \ddots & \vdots \\ 2\pi_{n_i1} & \cdots & 1 + d_{i_{n_i}} \end{pmatrix},$$

and the corresponding kinship coefficient matrix for both subject i and his/her relatives are defined by $\boldsymbol{\Psi_i}$. We denote a $w \times w$ dimensional identity matrix by I_w, and w dimensional column vector of which all elements are 0 and 1 by 0_w and 1_w respectively. If we let σ_g^2 and σ_ϵ^2 be variances of polygenic effect and random effect, and Z_i is assumed to include the intercept, we can assume that

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

Here we assume no main genetic effects (thus, $\boldsymbol{G}_{\mathrm{i}}$ is not included in the model) and $\sigma_{\epsilon}^{2} = 1$. If the heritability, h^2, and the prevalence of the disease, q are known, then liability threshold T and σ_g^2 can be evaluated from the standard normal distribution, i.e.,

$$\Phi\left(-\frac{T}{\sqrt{\sigma_g^2 + 1}}\right) = 1 - q.$$

Here we used $q = 0.099$. Under this liability threshold model, the conditional mean (CM) of disease risk can be defined as,

$$CM = E\left(\boldsymbol{L}_i \middle| \boldsymbol{Y}_{\boldsymbol{i}_j} = \boldsymbol{y}_{\boldsymbol{i}_j}\right) \tag{S1}$$

Note that neither direct genotyped variables nor environmental variables (which is not available for the relatives) were included in our liability model. However, it is straightforward to extend the model in the existence of those variables.

Calculation of the conditional mean

Based on this liability threshold model, we can calculate the conditional mean of L_i, or CM, when conditioning on family history. To make the phenotype in liability scale, let $\mathrm{I}_{L_{i_j}} = 1$ if $L_{i_j} \in (T, \infty)$ and 0 otherwise. Then $\mathrm{I}_{\boldsymbol{L_i}} = \left(\mathrm{I}_{L_i}, \ldots, \mathrm{I}_{L_{i_{n_i}}}\right)^t$ and the CM of for subject i becomes $\mathrm{E}\left(L_i \middle| \mathrm{I}_{\boldsymbol{L}_{i_j}} = \boldsymbol{1}_{n_i}\right)$, which can be calculated using the truncated multivariate normal distribution. Here and n_i dimensional column vector of which all elements are 1 by $\boldsymbol{1}_{n_i}$.

To calculate the CM, we assume a liability vector, $\boldsymbol{L}$ (hereafter we dropped the subscripts for simplicity) follows n-dimensional multivariate normal distribution and the probability density function (*pdf*) is,

$$f(\boldsymbol{L}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{L}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{L}\right) \tag{S2}$$

where $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{L})$. Correspondingly, the conditional *pdf* of $\boldsymbol{L}$ becomes

$$f_\alpha(\boldsymbol{L}) = \begin{cases} \frac{1}{\alpha} f(\boldsymbol{L}), & \text{for } (T, \infty) \\ 0, & \text{otherwise} \end{cases} \tag{S3}$$

where $\alpha = Pr(\boldsymbol{L} \in (T, \infty))$, the fraction after truncation.

The moment generating approach can be applied to the truncated density in Eq (3) and provides the *n*-dimensional truncated *mgf* (Manjunath and Wilhelm, 2012),

$$m(\boldsymbol{t}) = \frac{\exp\left(\frac{\boldsymbol{t}^t \boldsymbol{\Sigma} \boldsymbol{t}}{2}\right)}{\alpha (2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \int_A \exp\left(-\frac{1}{2} \boldsymbol{L}^t \boldsymbol{\Sigma} \boldsymbol{L}\right) d\boldsymbol{L} \tag{S4}$$

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

$$\mu_i = \frac{\partial m(\boldsymbol{t})}{\partial t_i}\Big|_{t=0} = \sum_{k=1}^{n} \sigma_{ik} F_k^* \tag{S5}$$

where

$$F_k^* = \begin{cases} F_k(T) - F_k(\infty) & \text{if } y_k = 1 \\ F_k(-\infty) - F_k(T) & \text{otherwise} \end{cases} \tag{S6}$$

The F_k can be similarly derived as was done in (Manjunath and Wilhelm, 2012). First, we partition $\boldsymbol{L}$ into two parts, L_i and $\boldsymbol{L}_{i_j}$, and then $\boldsymbol{L}$ can be rewritten as

$$\boldsymbol{L} = \begin{pmatrix} L_i \\ \boldsymbol{L}_{i_j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ \boldsymbol{0}_n \end{pmatrix}, \begin{pmatrix} 1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \tag{S7}$$

If we denote the lower and upper truncated bound of $\boldsymbol{L}$ respectively as $\boldsymbol{a}$ and $\boldsymbol{b}$, *i.e.*, $\boldsymbol{a} < \boldsymbol{L} < \boldsymbol{b}$ where $\boldsymbol{a} = \left(a_i, \boldsymbol{a}_{i_j}\right)^t$, $\boldsymbol{b} = \left(b_i, \boldsymbol{b}_{i_j}\right)^t$. Then the truncated normal distribution function is

$$f_\alpha\left(\boldsymbol{L}_{i_j}, L_i = x\right) = \alpha^{-1} f(L_i = x) f\left(\boldsymbol{L}_{i_j} | L_i = x\right) I_\alpha. \tag{S8}$$

Because a conditional distribution of a normal distribution is also normally distributed, $\boldsymbol{L}_{i_j}|L_i = x$ is normally distributed with $E\left(\boldsymbol{L}_{i_j}|L_i = x\right) = \boldsymbol{\Sigma}_{12}x$ and $Var\left(\boldsymbol{L}_{i_j}|L_i = x\right) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}^{t}$. Accordingly, the multivariate marginal pdf of $\boldsymbol{L}_{i_j}$ becomes

$$F_{\boldsymbol{L}_{i_j}}(x) = \int_{a_{i_j}}^{b_{i_j}} \alpha^{-1} f(L_i = x) f\left(\boldsymbol{L}_{i_j}|L_i = x\right) dL_{i_j}. \quad \text{(S8)}$$

The integral is over L_{i_j} and thus $F_{\boldsymbol{L}_{i_j}}(x)$ can be rewritten as $\alpha^{-1} f(L_i = x) \int_{a_{i_j}}^{b_{i_j}} f\left(\boldsymbol{L}_{i_j}|L_i = x\right) dL_{i_j}$, which can be readily computed by using conventional statistical software. For this purpose, we used the *pmvnorm()* function in *mtvnorm* R package.

Note that if we denote the number of relatives for subject i by n_i the convergence rate of the MC algorithm for subject i and for all subjects are $O(\sqrt{n_i})$ and $O(\sqrt{n_1} + \cdots + \sqrt{n_n})$, respectively (http://arvix.org/pdf/1206.5387.pdf). Thus, computational intensity is positively related with the number of relatives and subjects and if the same size is large and many relatives' phenotypes are known, it can be computationally intensive.

Supplementary Figures and Tables

Figure S1. MDS plot of two datasets.

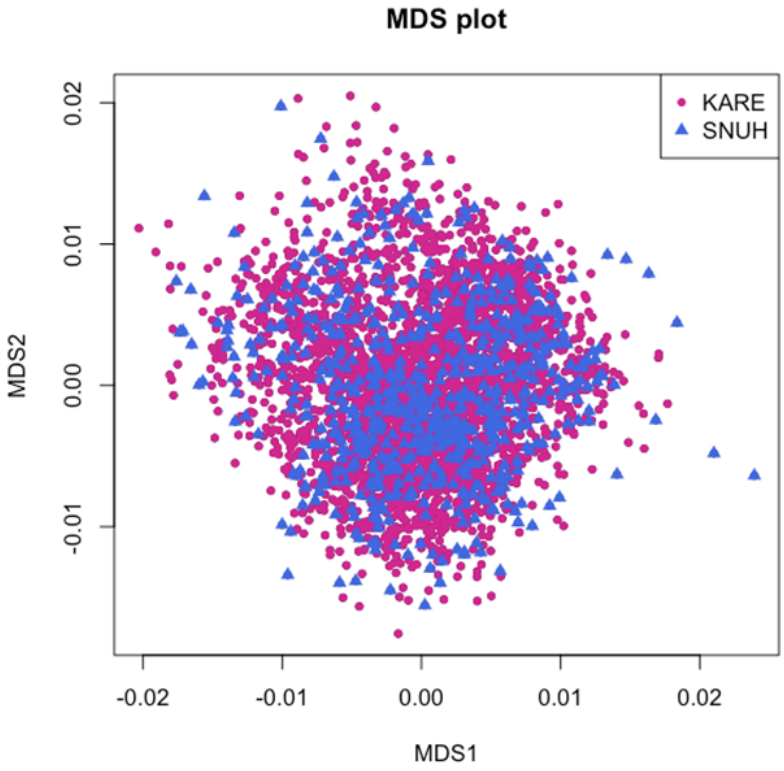

Figure S2. MAF scatter plot. SNPs MAFs (~300k) in SNUH and KARE datasets are plotted.

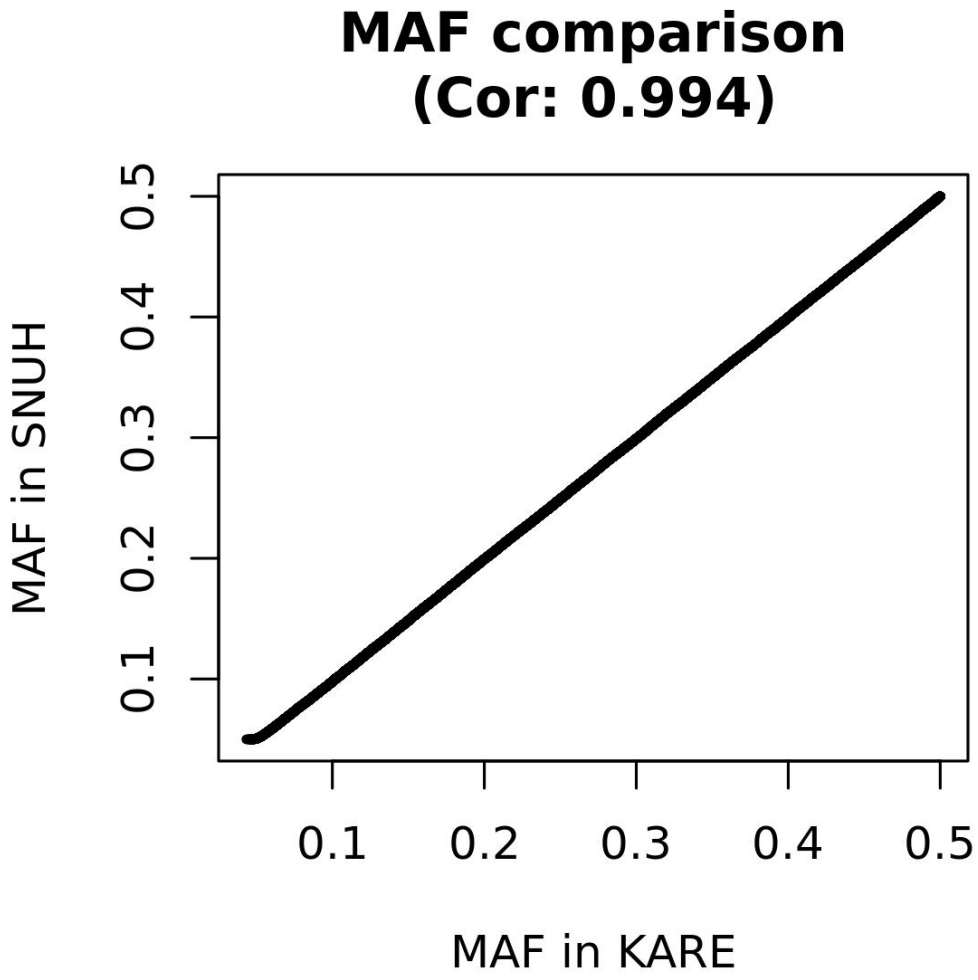

Figure S3. Characteristics of selected SNPs. To find the most effective set of SNPs, we selected SNPs based on the p-value obtained from the logistic regression and the BLUP obtained by the mixed effects model. Since the selected set of SNPs should be applied in penalized regression, we expected that the selection procedure would be more effective if the set of SNPs was uniformly distributed across the genome. Toward this end, we divided the whole genome into 3,233 windows of size 5M and counted the frequency of SNPs in each window. With a varying number of SNPs (0.1k – 20k).

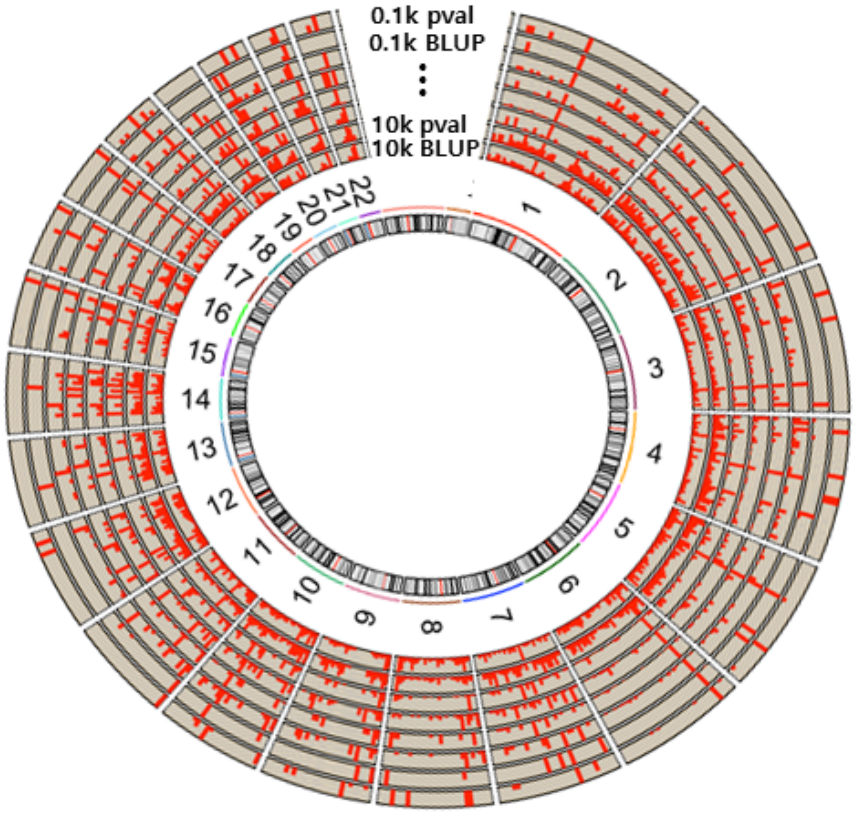

Figure S4. Proportion of variation explained by each variable in the final model.

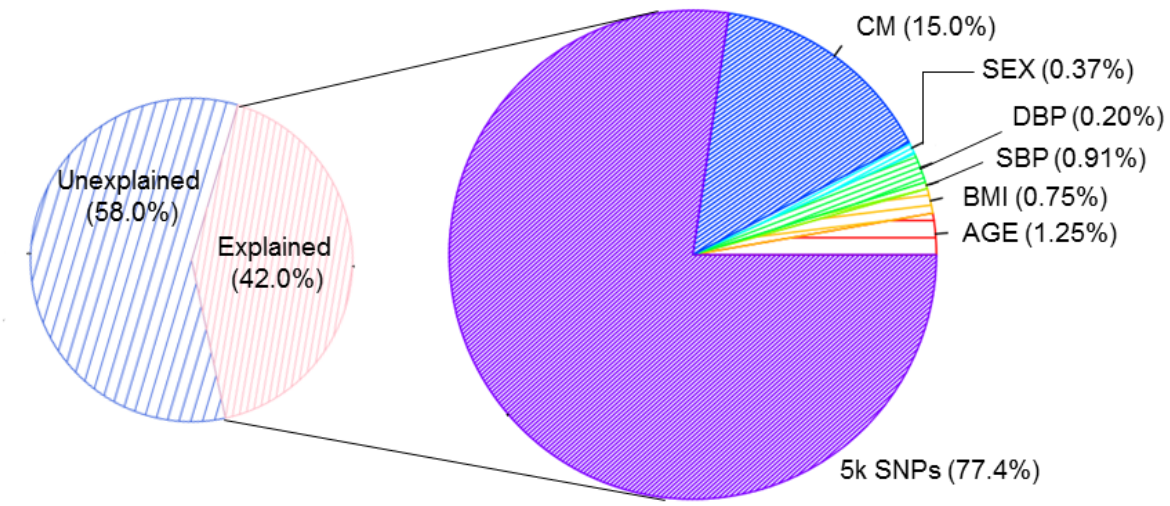

Figure S5. Proportion of variation explained by each variable in the final model without CM variable. For five clinical variables (age, sex, BMI, SBP, DBP) except CM variable, the individual proportions of the variation are shown, whereas the variation explained by the 5,000 SNPs is shown according to their summed proportion.

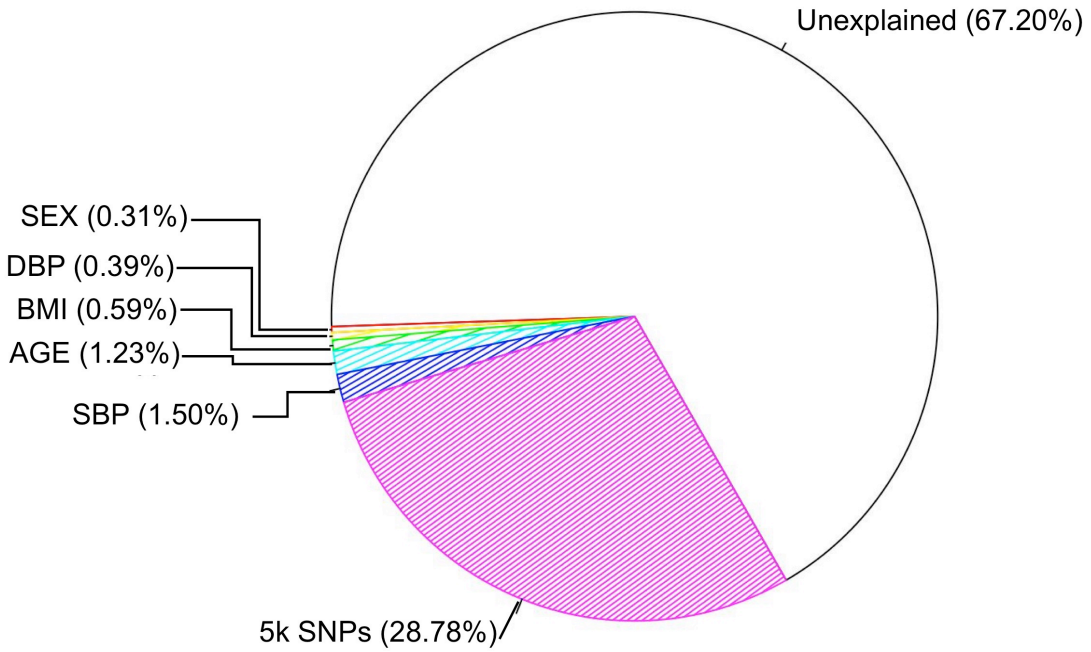

Table S1 Model comparison with different family history measures (AUC)

Family History	# of SNPs	RIDGE	LASSO	EN	SCAD	T.RIDGE
None	100	0.611 (0.023)	0.602 (0.025)	0.602 (0.023)	0.585 (0.023)	0.612 (0.024)
	500	0.614 (0.029)	0.600 (0.031)	0.600 (0.028)	0.594 (0.031)	0.614 (0.030)
	1,000	0.626 (0.029)	0.611 (0.029)	0.611 (0.031)	0.601 (0.032)	0.628 (0.029)
	5,000	0.689 (0.032)	0.647 (0.030)	0.647 (0.030)	-	0.689 (0.031)
	10,000	0.672 (0.028)	0.626 (0.029)	0.626 (0.030)	-	0.672 (0.030)
	20,000	0.674 (0.030)	0.639 (0.031)	0.639 (0.029)	-	0.674 (0.031
Weighted Mean	100	0.669 (0.028)	0.605 (0.030)	0.605 (0.030)	-	0.615 (0.030)
	500	0.617 (0.028)	0.602 (0.026)	0.602 (0.026)	-	0.617 (0.020)
	1,000	0.629 (0.028)	0.615 (0.026)	0.615 (0.026)	-	0.630 (0.031)
	5,000	0.692 (0.028)	0.650 (0.024)	0.650 (0.024)	-	0.692 (0.012)
	10,000	0.676 (0.032)	0.630 (0.035)	0.630 (0.035)	-	0.676 (0.012)
	20,000	0.683 (0.033)	0.647 (0.037)	0.647 (0.037)	-	0.683 (0.012)
Conditional Mean	100	0.669 (0.023)	0.659 (0.024)	0.659 (0.024)	0.643 (0.021)	0.669 (0.021)
	500	0.659 (0.030)	0.642 (0.031)	0.642 (0.031)	0.639 (0.030)	0.659 (0.031)
	1,000	0.670 (0.029)	0.651 (0.029)	0.651 (0.029)	0.645 (0.028)	0.670 (0.030)
	5,000	0.736 (0.030)	0.691 (0.031)	0.691 (0.031)	-	0.736 (0.027)
	10,000	0.721 (0.029)	0.673 (0.030)	0.673 (0.029)	-	0.721 (0.031)
	20,000	0.725 (0.034)	0.689 (0.031)	0.689 (0.031)	-	0.725 (0.030)

Table S2 Model comparison with different SNP filtering criteria (without CM variable)

CRITERIA	# of SNPs	RIDGE (SD)	LASSO (SD)	EN (SD)	SCAD (SD)	T.RIDGE (SD)
P-value	100	0.642 (0.024)	0.637 (0.022)	0.637 (0.022)	0.616 (0.026)	0.641 (0.021)
	500	0.640 (0.032)	0.626 (0.031)	0.626 (0.031)	0.608 (0.032)	0.640 (0.031)
	1,000	0.640 (0.027)	0.624 (0.028)	0.624 (0.028)	0.608 (0.029)	0.640 (0.026)
	5,000	0.660 (0.029)	0.635 (0.029)	0.635 (0.029)	-	0.660 (0.027)
	10,000	0.668 (0.028)	0.640 (0.030)	0.640 (0.030)	-	0.668 (0.030)
	20,000	0.674 (0.031)	0.640 (0.026)	0.640 (0.026)	-	0.674 (0.028)
BLUP	100	0.611 (0.023)	0.602 (0.025)	0.602 (0.023)	0.585 (0.023)	0.612 (0.024)
	500	0.614 (0.029)	0.600 (0.031)	0.600 (0.028)	0.594 (0.031)	0.614 (0.030)
	1,000	0.626 (0.029)	0.611 (0.029)	0.611 (0.031)	0.601 (0.032)	0.626 (0.029)
	5,000	0.689 (0.032)	0.647 (0.030)	0.647 (0.031)	-	0.689 (0.031)
	10,000	0.672 (0.028)	0.626 (0.029)	0.626 (0.030)	-	0.672 (0.030)
	20,000	0.674 (0.030)	0.639 (0.031)	0.639 (0.029)	-	0.674 (0.031)

Table S3 Model comparison with different SNP filtering criteria (with CM variable)

CRITERIA	# of SNPs	RIDGE (SD)	LASSO (SD)	EN (SD)	SCAD (SD)	T.RIDGE (SD)
P-value	100	0.693 (0.025)	0.687 (0.024)	0.688 (0.023)	0.676 (0.025)	0.693 (0.023)
	500	0.687 (0.031)	0.672 (0.030)	0.672 (0.031)	0.665 (0.032)	0.687 (0.034)
	1,000	0.685 (0.028)	0.669 (0.029)	0.669 (0.031)	0.664 (0.029)	0.685 (0.029)
	5,000	0.709 (0.030)	0.687 (0.031)	0.687 (0.027)	-	0.709 (0.031)
	10,000	0.717 (0.029)	0.690 (0.027)	0.690 (0.029)	-	0.717 (0.028)
	20,000	0.721 (0.030)	0.689 (0.031)	0.689 (0.027)	-	0.721 (0.030)
BLUP	100	0.669 (0.023)	0.659 (0.024)	0.659 (0.024)	0.643 (0.021)	0.669 (0.021)
	500	0.659 (0.030)	0.642 (0.031)	0.642 (0.031)	0.639 (0.030)	0.659 (0.031)
	1,000	0.670 (0.029)	0.651 (0.029)	0.651 (0.029)	0.645 (0.028)	0.670 (0.030)
	5,000	**0.736** (0.030)	0.691 (0.031)	0.691 (0.031)	-	**0.736** (0.027)
	10,000	0.721 (0.029)	0.673 (0.030)	0.673 (0.029)	-	0.721 (0.031)
	20,000	0.725 (0.034)	0.689 (0.031)	0.689 (0.032)	-	0.725 (0.030)