Bioinformatics, 36(22-23), 2020, 5415–5423
doi: 10.1093/bioinformatics/btaa1023
Advance Access Publication Date: 17 December 2020
Original Paper

OXFORD

Genetics and population analysis

A Bayesian linear mixed model for prediction of complex traits

Yang Hai and Yalu Wen *

Department of Statistics, University of Auckland, Auckland 1010, New Zealand

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on July 26, 2020; revised on November 24, 2020; editorial decision on November 25, 2020; accepted on November 27, 2020

Abstract

Motivation: Accurate disease risk prediction is essential for precision medicine. Existing models either assume that diseases are caused by groups of predictors with small-to-moderate effects or a few isolated predictors with large effects. Their performance can be sensitive to the underlying disease mechanisms, which are usually unknown in advance.

Results: We developed a Bayesian linear mixed model (BLMM), where genetic effects were modelled using a hybrid of the sparsity regression and linear mixed model with multiple random effects. The parameters in BLMM were inferred through a computationally efficient variational Bayes algorithm. The proposed method can resemble the shape of the true effect size distributions, captures the predictive effects from both common and rare variants, and is robust against various disease models. Through extensive simulations and the application to a whole-genome sequencing dataset obtained from the Alzheimer's Disease Neuroimaging Initiatives, we have demonstrated that BLMM has better prediction performance than existing methods and can detect variables and/or genetic regions that are predictive.

Availabilityand implementation: The R-package is available at https://github.com/yhai943/BLMM.

Contact: y.wen@auckland.ac.nz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The concept of treating diseases with precise interventions, designed rationally from a detailed understanding of the biological mechanisms and individual differences, has been widely accepted as the goal of precision medicine, an emerging model of healthcare that tailors treatment strategies based on individuals' profiles (Ashley, 2015). Towards this end, there is an expectation that the emerging genetic findings and other existing knowledge will revolutionize the current trial-and-error practice of medicine by enabling more accurate disease prediction and precise prevention/treatment strategies. It is expected that the formed risk prediction models can help identify high-risk sub-populations so that appropriate treatments can be delivered. Despite promising, for most of the complex human diseases, the existing models lack sufficient accuracy for clinical use (Lipinski *et al.*, 2016).

Human diseases are usually affected by multiple genetic variants through complex biological mechanisms, and thus progress towards accurately predicting disease risk requires the development of analytical models that can jointly consider multiple genetic variants and allow for various magnitudes of effect sizes for different predictors. The genomic best linear unbiased prediction (gBLUP) models have long been used for risk prediction research (Henderson, 1950; Moser *et al.*, 2009). They use the linear mixed model (LMM) framework to model disease risk, where high-dimensional genomic data is jointly modelled through the variance-covariance matrix of the random effect. Instead of modelling the effects from individual predictors, they aim at estimating the cumulative predictive effects from multiple genetic variants, making it possible to simultaneously consider a large amount of predictors (De los Campos *et al.*, 2013; Hayes *et al.*, 2001). For example, Yang *et al.* predicted human heights using the gBLUP method, where the cumulative predictive effect of all genetic variants is estimated using the LMM framework with a single random effect term. Their model explains 45% heritability that is considerably higher than the model built with known predictors (Yang *et al.*, 2010). Although practically useful and easy to implement, gBLUP simply assumes all genetic variants have the same effect-size distribution (Henderson, 1975), which is too simple to be realistic (Speed and Balding, 2014). For example, single nucleotide polymorphisms (SNPs) that come from different genetic regions (e.g. coding and intron SNPs) are unlikely to have the same type of effect sizes (Speed and Balding, 2014). Over the past decades, many efforts have been made to relax the assumptions of gBLUP(Speed and Balding, 2014; Weissbrod *et al.*, 2016; Zeng and

Zhou, 2017; Zhou *et al.*, 2013). For example, MultiBLUP (Speed and Balding, 2014) splits the genome into multiple genetic regions and allows for variants from different regions having different effect size distributions. However, these extensions still aim at modelling the cumulative predictive effects from a group of predictors and thus may fail to capture the effects from isolated predictors (i.e. a very small fraction of variants across the genome and they are not located nearby).

While jointly modelling all genetic variants has great potential to capture predictive markers, the large amount of noise in high-dimensional genomic data can substantially attenuate the robustness and accuracy of prediction models. It is noted by Byrnes *et al.* (2013) that variable selection is of great importance for building risk prediction models in the absence of biological annotations. Sparsity regression model has been widely used in genetic research (Carvalho *et al.*, 2008; Zhou *et al.*, 2013). For example, in order to identify causal variants, a penalized maximum likelihood approach that introduces the normal exponential gamma density as a penalty function was developed, and the corresponding non-zero coefficients were used for prediction (Hoggart *et al.*, 2008). Logistic regression with L_1 penalty, which can shrink the coefficients of non-relevant variables towards zero, has been developed to select the most predictive genetic variants for cancer classification (Algamal and Lee, 2015). Different from the key assumption used in gBLUP that assumes genetic effects follow a normal distribution, the sparsity regression models assume that only a small fraction of genetic variants have moderate-to-large effect sizes and the rest are noise (Zhou *et al.*, 2013). Their performance can be affected by the underlying disease model that is usually unknown in advance (Chatterjee *et al.*, 2016). In addition, simultaneously modelling and selecting predictive variants from millions of potential predictors in the sparsity regression models can be computationally challenging.

Bayesian LMMs are another widely used alternatives for the analysis of genomic data (Dunson, 2001), and they have shown better prediction performance than gBLUP-based methods (Zeng and Zhou, 2017; Zhou *et al.*, 2013). Compared to the frequentest, Bayesian LMMs can easily accommodate various model assumptions and different types of effects via specifying different prior distributions (Zhao *et al.*, 2006). For example, BayesA assumes each genetic effect has its own variance and thus a scaled univariate student's t prior distribution (i.e. $\beta_i \sim t(0, t, \sigma_\alpha^2)$) is used (Habier *et al.*, 2011; Zhou *et al.*, 2013). BayeB assumes that a large amount of genetic variants have no predictive effects, and thus a mixture distribution ($\beta_i \sim \pi t(0, t, \sigma_\alpha^2) + (1-\pi)\delta_0,\ \sigma_\alpha^2 \neq 0$) that uses a t-distribution to account for predictive effects and a point mass at zero to model the effects from noise variants is set as it is prior (Habier *et al.*, 2011; Zhou *et al.*, 2013). Bayesian Lasso assumes the majority of variants have weak or no effects, and thus the double exponential (DE) prior distribution ($\beta_i \sim DE(0, \theta)$) is employed (Yi and Xu, 2008). Existing Bayesian LMMs usually assume that only a very small fraction of genetic variants are predictive and set the priors accordingly. However, their performance can be sensitive to the underlying disease model (Zeng and Zhou, 2017; Zhou *et al.*, 2013) and the choice of priors adopted (Gianola, 2013). While previous attempts mainly focus on modifying the parametric priors so that they can mimic the distribution of true effect sizes, none of them work uniformly better as each disease can have its own mechanism and the true effect distribution is usually unknown. Recently, a non-parametric Bayesian method, latent Dirichlet process regression (DPR), was proposed to better model the effect size distributions. It uses the Dirichlet process normal mixture (i.e. $\beta_i \sim \pi_1 N(0, 0 \times \sigma_e^2) + \sum_{k=2}^{\infty} \pi_k N(0, \sigma_k^2 \sigma_e^2)$) as its prior for each genetic effect, and thus has the flexibility to model any unknown distributions (Zeng and Zhou, 2017). However, it cannot consider the existing biological annotations and thus fails to capture small effects from multiple genetic variants located within a functional unit (e.g. gene).

gBLUP-based methods and their Bayesian counterparts have made various levels of successes in risk prediction research (Yang *et al.*, 2010; Zeng and Zhou, 2017; Zhou *et al.*, 2013). However, there are several key limitations. First, the existing methods are mainly designed for common variants (i.e. minor allele frequency > 5%) and thus are not capable of capturing the predictive effects from rare variants. Converging evidences have suggested that rare variants can play an important role in explaining disease heritability (Gibson, 2012). For example, Bodmer and Bonilla (2008) found rare variants in *APC* gene are associated with colorectal cancer. Overlooking the contribution of rare variants can substantially reduce the accuracy of risk prediction models (Eichler *et al.*, 2010). Second, most of the existing methods assume that diseases are either caused by a large number of predictors with small-to-moderate effects (e.g. gBLUP) or a few isolate predictors with large effects (e.g. sparsity regression). Therefore, their performance can be sensitive to the underlying disease model. Third, identifying predictors from high-dimensional data is a daunting task (Hoggart *et al.*, 2008; Yi and Xu, 2008; Zhou *et al.*, 2013), especially when a large amount of rare variants are considered. While it is widely accepted that jointly selecting predictors located in nearby regions (e.g. all markers within a gene) can improve computational efficiency and the interpretation of the prediction model, there lack theoretical support and empirical criteria are often used (Speed and Balding, 2014; Weissbrod *et al.*, 2016).

To address these challenges, we developed a Bayesian LMM with multiple random effects (denoted as BLMM). The proposed model can (i) resemble the true effect size distributions, (ii) capture the predictive effects from both common and rare variants and (iii) is robust against disease models. It can simultaneously select isolated predictors with large effects and a group of predictors with small-to-large effects. Rather than using Markov chain Monte Carlo (MCMC) that can be intractable, we developed a variational Bayesian (VB) algorithm that provides an analytical approximation to the posterior distribution (Blei *et al.*, 2017). In the following sections, we first presented the BLMM model and the variational distribution for its parameter estimation. We further compared its predictive ability with existing methods, including (i) MultiBLUP (Speed and Balding, 2014) and (ii) DPR (Zeng and Zhou, 2017). Finally, we illustrated the advantage of BLMM with an application to the dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller *et al.*, 2005).

2 Materials and methods

In the following sections, we first presented the prediction model and then provided a detailed description of the VB algorithm used for the parameter estimation.

2.1 BLMM for risk prediction of complex traits

It is well accepted that the predictive effects of genetic variants can differ substantially across genomic regions (e.g. intergenic, Exon and Intron) (Schork *et al.*, 2013; Speed and Balding, 2014), and some predictors can have very strong effects (e.g. the risk allele of *APOE* gene on predicting Alzheimer's Disease). Therefore, given M genetic regions (e.g. genes), we proposed to form a hybrid model as

$$\boldsymbol{Y} = \boldsymbol{\beta} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \boldsymbol{g}_{\mathrm{m}} + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_\epsilon^2), \tag{1}$$

where $\boldsymbol{Y}$ is the outcome and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_\epsilon^2)$. Similar to existing Bayesian methods, we set the prior for σ_ϵ^2 as $\sigma_\epsilon^2 \sim IG(a_0, b_0)$, where IG denotes the inverse gamma distribution and $a_0 = b_0 = 0.1$. $\boldsymbol{X}(\boldsymbol{\beta})$ is the genotypes (their effects) and they are designed to capture the isolated large predictive effects. $\boldsymbol{g}_m$, on the other hand, represents the cumulative predictive effects from all genetic variants within region m, and they are used to reflect the effects from a group of predictors. The rationale of having $\boldsymbol{\beta}$ in the proposed model is similar to that used in the sparsity regression models, which assume disease heritability can be mainly explained by a few isolated genetic markers with large effects (Yi and Xu, 2008; Zhou *et al.*, 2013). The rationale of having $\boldsymbol{g}_m$ in the proposed model is similar to those

adopted in gBLUP and their extensions, which assume disease heritability can be attributed to a large number of genetic markers with small-to-moderate effects (Speed and Balding, 2014; Weissbrod *et al.*, 2016; Zhou *et al.*, 2013). In addition, similar to the random field model proposed by Wen *et al.* (2016), $\boldsymbol{g}_m$ also has the potential to capture the cumulative predictive effect from all rare variants within region m. Therefore, the proposed BLMM is a very flexible framework for modelling diseases with various underlying mechanisms.

2.1.1 The predictive effects from isolated predictors

To detect the isolated predictors with large effects (i.e. $\boldsymbol{\beta}$), the spike and slab prior can be used for each β_j, where the effects from noise markers can have a point mass at zero. However, such a procedure can underestimate the posterior variances for $\boldsymbol{\beta}$ (Carbonetto *et al.*, 2012). To address this issue, we introduced a Bernoulli random vector and re-parameterized the model as,

$$\boldsymbol{Y} = \boldsymbol{X\Gamma\beta} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \boldsymbol{g}_{\mathrm{m}} + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}\sigma_\epsilon^2), \tag{2}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{\mathrm{p}})^{\mathrm{T}}$ is a vector of binary variables indicating whether each genetic variant is included in the model. We set the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as $\beta_j \sim \mathcal{N}(0, \sigma_\beta^2)$ and $\gamma_j \sim \mathrm{Bernoulli}(\theta_0)$, respectively. We used $\theta_0 \in [0, 1]$ as a tuning parameter to control the sparsity level. Although the principle of indifference usually sets $\theta_0 = 0.5$ (Cerquides and de Mantaras, 2003), we set $\theta_0 = 0.1$ for the proposed method as the probability for each variant to be predictive is much lower than 0.5 for high-dimensional genetic data.

Given model 2, it is straightforward to see that the genetic variant has no effect on the outcome when $\gamma_j = 0$, regardless of the value of β_j (i.e. $\beta_j \gamma_j X_j = 0$). By using $\boldsymbol{\gamma}$ to perform variable selection, the proposed re-parametrization can keep all of β_j in the same partition and thus avoids the underestimation of their posterior variances. It can be shown that the isolated genetic effects in the proposed model follow the Bernoulli-Gaussian distribution(Fernandes *et al.*, 2017; Ormerod *et al.*, 2017). The details of the derivation of conditional posterior can be found in Appendix SA.1.

2.1.2 Cumulative predictive effects from groups of predictors

To capture the cumulative predictive effects (i.e. $\boldsymbol{g}_m$) from a group of predictors located nearby, we followed a similar idea used in MultiBLUP (Speed and Balding, 2014) and assumed that genetic similarities can lead to phenotypic similarity. We allow different regions contributing differently to the outcome, and set a multivariate normal prior for each region-based cumulative predictive effect (i.e. $\boldsymbol{g}_m$) as

$$\begin{aligned} \boldsymbol{g}_m | \boldsymbol{K}_m &\sim \mathcal{N}(0, \boldsymbol{K}_m \sigma_m^2) \quad m = 1, \ldots, M \\ \sigma_m^2 &\sim IG(a_1, b_1). \end{aligned} \tag{3}$$

$\boldsymbol{K}_m$ is the genetic similarity for region m and it is defined as $\boldsymbol{K}_m = \boldsymbol{G}_m \boldsymbol{W}_m \boldsymbol{G}_m^T / p_m$, where $\boldsymbol{G}_m$ is the genotype matrix for region m and p_m is the number of genetic markers in the region. $\boldsymbol{W}_m = \mathrm{diag}(w_1, w_2, \ldots, w_{p_m})$ is the pre-specified weights used to capture the contribution of rare variants. Similar to existing literature (Wu *et al.*, 2011), we defined the weights as $w_j = \frac{1}{MAF_j(1-MAF_j)}$, where MAF_j is the minor allele frequency for the jth variant. σ_m^2 reflects the effect sizes for predictors in region m, and they allow to differ across different regions (i.e. $\sigma_1^2 = \sigma_2^2 \cdots = \sigma_M^2$ is not required). Since the predictive effects from a group of predictors are usually assumed to be small-to-moderate except for a few rare variants (Saint Pierre and Genin, 2014; Zhou *et al.*, 2013), the values of σ_m^2 are expected to be small. Therefore, the hyper-parameters (i.e. a_1 and b_1) are set to be 0.1 for all regions.

For high-dimensional data, many regions included in the analysis are not disease-related, and these noise regions can reduce the robustness and accuracy of the prediction models. Moreover, the identification of genetic susceptibility regions facilitate the model interpretation and help to determine sub-populations at high risk (Weissfeld *et al.*, 2015). For the proposed model, the identification of predictive regions is equivalent to determine which $\sigma_m^2 \neq 0$. Therefore, rather than assuming all regions have predictive effects (i.e. Equation 3), we propose to use the idea of the spike and slap prior (Mitchell and Beauchamp, 1988) and set the prior for each region as

$$\begin{aligned} \boldsymbol{g}_m | \boldsymbol{K}_m, \sigma_m^2 &\sim \wp(r_m) N(0, \boldsymbol{K}_m \sigma_m^2) + (1 - \wp(r_m)) \delta_0, \quad m = 1, \ldots, M \\ r_m &\sim \mathrm{Bernoulli}(\theta_1), \end{aligned} \tag{4}$$

where r_m is a binary random variable indicating whether each genetic region is included in the model, $\wp(r_m)$ is the probability of success for a Bernoulli random trial r_m, and δ_0 denotes a discrete measure concentrated at zero. Since the majority of genetic regions explain little phenotypic variance, we set $\theta_1 = 0.1$ for the hyper-parameter in the Bernoulli distribution.

Directly incorporating Equation 4 into the proposed model is computationally demanding. To expedite its computation, we used the idea from Chen and Dunson (2003) and decomposed $\boldsymbol{K}_m$ as $\boldsymbol{K}_m = \boldsymbol{Q}_m \boldsymbol{\Lambda}_{\mathrm{m}} \boldsymbol{Q}_m^T$, where $\boldsymbol{\Lambda}_{\mathrm{m}} = diag(\lambda_{m1}, \ldots, \lambda_{mn})$ with $\lambda_{m1} \geq \lambda_{m2} \cdots \geq \lambda_{mn} \geq 0$ being eigenvalues and $\boldsymbol{Q}_m$ is a matrix of the corresponding eigenvectors. We further re-parameterized the cumulative predictive effects part with the slab and spike prior as

$$\boldsymbol{Y} = \boldsymbol{X\Gamma\beta} + \sum_{\mathrm{m}=1}^{\mathrm{M}} (\boldsymbol{Z}_{\mathrm{m}} \mathrm{r}_{\mathrm{m}} U_{\mathrm{m}}) + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{U}_m \sim N(0, \boldsymbol{I}\sigma_m^2)$, $\boldsymbol{Z}_m = \boldsymbol{Q}_m \boldsymbol{\Lambda}_{\mathrm{m}}^{\frac{1}{2}}$ and $E(r_m = 1) = \wp(r_m)$. While the re-parameterization facilitates the selection of predictive regions (i.e. $r_m = 1$ indicates the region is predictive), the dimensions of both $\boldsymbol{Z}_m$ and $\boldsymbol{U}_m$ depend on the sample size n, making their manipulations still computationally challenging. Note that $\boldsymbol{K}_m$ is only guaranteed to be positive semi-definite and the eigenvalues decay very fast. Therefore, we used the idea of kernel principal component analysis and approximated $\boldsymbol{Z}_m$ with a low-rank matrix $\boldsymbol{Z}'_m = \boldsymbol{Q}'_m (\boldsymbol{\Lambda}'_{\mathrm{m}})^{\frac{1}{2}}$, where $\boldsymbol{\Lambda}'_{\mathrm{m}}$ is a diagonal matrix with the n_m largest eigenvalues and $\boldsymbol{Q}'_m$ is the corresponding eigenvectors. Equation 5 can be written as

$$\boldsymbol{Y} = \boldsymbol{X\Gamma\beta} + \sum_{\mathrm{m}=1}^{\mathrm{M}} (\boldsymbol{Z}'_{\mathrm{m}} \mathrm{r}_{\mathrm{m}} U'_{\mathrm{m}}) + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{U}'_m \sim N(0, \boldsymbol{I}'\sigma_m^2)$ with $\boldsymbol{I}'$ being an $n_m \times n_m$ identity matrix.

2.2 Parameter estimation based on a variational Bayes algorithm

Estimating parameters for the proposed model using the standard techniques (e.g. MCMC) can be intractable and slow for convergence. Therefore, we developed a mean-field VB algorithm to infer parameters. Unlike the exact inference, VB approximates the posterior through minimizing the Kullback-Leibler (KL) divergence between the true and approximation distributions (Zhang *et al.*, 2019), and it converges much faster than the exact method (Salimans *et al.*, 2015; Zhang *et al.*, 2019).

The basic idea of VB is to find a family of simple probability distributions that approximate the true posterior distribution as close as possible in terms of KL divergence, and the parameters in the approximated distributions can be easily estimated (Ghahramani and Beal, 2000; Zhang *et al.*, 2019). Let $\boldsymbol{\xi}$ be all the parameters that need to be estimated. To choose a class of distribution that leads to a more traceable approximation (Bishop, 2006), we factorized the posterior $p(\boldsymbol{\xi}|\boldsymbol{y})$ as a product of independent distributions on small subsets of parameters. We defined the approximated distribution $q(\boldsymbol{\xi})$ as $q(\boldsymbol{\xi}) = \mathrm{q}_{\boldsymbol{\beta}} \prod_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{q}_{\gamma_{\mathrm{j}}} \prod_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{q}_{U_{\mathrm{m}}} \prod_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{q}_{\mathrm{r}_{\mathrm{m}}} \prod_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{q}_{\sigma_{\mathrm{m}}^2} \cdot \mathrm{q}_{\sigma_\epsilon^2}$, where $q_{\boldsymbol{\beta}} = \mathcal{N}(\boldsymbol{M}_{\boldsymbol{\beta}}, \boldsymbol{S}_{\boldsymbol{\beta}})$; $q_{\gamma_j} = \mathrm{Bernoulli}(w_j)$; $q_{U_m} = \mathcal{N}(\boldsymbol{M}_m, \boldsymbol{S}_m)$; $q_{r_m} = \mathrm{Bernoulli}(w_m)$; $q_{\sigma_m^2} = IG(a_m, b_m)$; and $q_{\sigma_\epsilon^2} = IG(a_\epsilon, b_\epsilon)$. We estimated the parameters by minimizing the KL divergence between the exact posterior distribution ($p(\boldsymbol{\xi}|\boldsymbol{y})$) and the variational distribution $q(\boldsymbol{\xi})$:

Algorithm 1: Inference procedure of variational Bayes algorithm for BLMM.

Input: X, G_1, ..., G_M, y
Output: $\hat{\beta}$, w1, ..., wp, $\hat{U}_1$, ..., $\hat{U}_M$, w1, ..., wM, $\hat{\sigma}_1^2$, ..., $\hat{\sigma}_M^2, \hat{\sigma}_\epsilon^2$
Initialisation: define $K_m \propto G_m W_m G_m^T$ for each region and set $Z_m = Q_m \Lambda_m^{\frac{1}{2}}$, m = 1 ..., M;
while the increase of variational lower bound (ELBO) is not negligible **do**
1: For individual effects: a). update M_β and S_β for β (see 9), and b) update $logit(w_1), \ldots, logit(w_p)$ (see 10);
2: For cumulative effects: a). update M_m and S_m for U_m (see 11); b) update $logit(w_1), \ldots, logit(w_M)$; see 12; and c) update a_m and b_m for σm (see 13);
3: Update a_ϵ and b_ϵ for σ_ϵ^2 according to 14;
end

$$KL(q|p)=\int q(\xi)\log\frac{q(\xi)}{p(\xi|y)}d\xi=\log p(y)-(E_{q(\xi)}[log p(\xi,y)]-E_{q(\xi)}[log q(\xi)]) \tag{7}$$

Minimizing $KL(q|p)$ is equivalent to maximizing the evidence lower bound (ELBO) calculated as

$$\begin{aligned}\text{ELBO} = {} & E_{q(\xi)}[\log p(\xi,y)] - E_{q(\xi)}[\log q(\xi)] = \log(\Gamma(a_\epsilon)) \\ & -a_\epsilon \log(b_\epsilon) + \frac{1}{2}\log(det(S_\beta)) \\ & + \frac{1}{2\sigma_\beta^2}(M_\beta T M_\beta + tr(S_\beta)) + \sum_j (w_j \log\frac{\log\theta_0}{w_j} \\ & + (1-w_j)\log\frac{1-\log\theta_0}{1-w_j}) + \sum_m (\log(\Gamma(a_m)) - a_m\log(b_m) \\ & + \frac{1}{2}\log(det(S_m)) - \frac{a_m}{2b_m}\left(M_m^T M_m + tr(S_m)\right) + w_m \log\frac{\theta_1}{w_m} \\ & + (1-w_m)\log\frac{1-\log\theta_1}{1-w_m}\Big),\end{aligned} \tag{8}$$

where $\psi(\cdot)$ is digamma function and $\Gamma(\cdot)$ is the gamma function; $w_j = p(\gamma_j = 1|y)$ and $w_m = p(r_m = 1|y)$. The details of ELBO derivations are shown in Appendix SA.2. To maximize the ELBO, we followed the same procedure used in Zhang *et al.* (2019), where parameters were updated one at a time using the coordinate ascent algorithm (Bishop, 2006). The inference procedure of the VB is shown in Algorithm 1. We used the estimating equations for updating parameter ξ as follows (details of derivation were shown in Appendix SA.3):

Update β**:** Recall the variation distribution for $q(\beta)$ is $q(\beta) = \mathcal{N}(M_\beta, S_\beta)$. The parameters M_β and S_β are updated according to estimating equations:

$$\begin{aligned} M_\beta &= E\left(\frac{1}{\sigma_\epsilon^2}\right) S_\beta E(\gamma_j) X^T E(A) \\ S_\beta &= \left\{ aE\left(\tfrac{1}{\sigma_\epsilon^2}\right)((X^T X)^\circ \Omega) + \mathrm{E}\left(\tfrac{1}{\sigma_\beta^2} I\right)\right\}^{-1}\end{aligned}$$

where $A = y - \sum_{m=1}^{M}(Z_m(r_m I_m U_m))$; $\Omega = ww^{\mathrm{T}} + W^\circ(I - W)$; $W = diag(w)$; $w = E_q(\gamma)$; and $^\circ$ denotes the Hadamard product. *Update* γ_j: The variation distribution for $q(\gamma_j)$ is $q(\gamma_j) = Bernoulli(w_j)$, where $w_j = E(q(\gamma_j))$. To update the γ_j, we used the following estimating equation (Ormerod *et al.*, 2017):

$$\begin{aligned} logit(w_j) = {} & logit(\theta_0) - \frac{1}{2}E\left(\frac{1}{\sigma_\epsilon^2}\right) X_j^T X_j (M_{\beta\beta(j)}^T M_{\beta(j)} + S_{\beta_{(j,j)}}) \\ & - E\left(\frac{1}{\sigma_{\epsilon^2}}\right) X_j^T \left(X_{(-j)}\Gamma_{(-j)}(M_{\beta(-j)} M_{\beta(j)} + S_{\beta_{(-j,j)}})\right) \\ & + E\left(\frac{1}{\sigma_{\epsilon^2}}\right) X_j^T (E(A) M_{\beta(j)}),\end{aligned}$$

where $logit(w_j) = \log\left(\frac{w_j}{1-w_j}\right)$; $M_{\beta(j)}$ is the *jth* component of M_β; $M_{\beta(-j)}$ is the whole vector of M_β except the single component at index j; $X_{(-j)}$ and $\Gamma_{(-j)}$ are the matrices of X and Γ except the jth column; and $S_{\beta_{(-j,j)}}$ is jth column of S_β without the *jth* component.

Update U_m: The variational distribution for $q(U_m)$ is $q(U_m) = \mathcal{N}(M_m, S_m)$, with mean M_m and variance S_m. Therefore, the parameters are updated as:

$$\begin{aligned} M_m &= E(\frac{1}{\sigma_{\epsilon^2}})\, S_m (E(r_m) I) Z_m^T E(B_m) \\ S_m &= \left\{ aE\left(\frac{1}{\sigma_{\epsilon^2}}\right)((Z_m^T Z_m)^\circ (w_m I_m)) + E\left(\frac{1}{\sigma_m^2} I\right)\right\}^{-1},\end{aligned} \tag{11}$$

where $B_m = y - X\Gamma\beta - \sum_{i\neq m}(Z_i(r_i I_i U_i))$.

Update r_m: The variation distribution for $q(r_m)$ is $q(r_m) = Bernoulli(w_m)$, where $w_m = E(q(r_m))$. Hence, the estimating equation for updating r_m is

$$\begin{aligned} logit(w_m) = {} & logit(\theta_1) - \frac{1}{2\sigma_\epsilon^2}\left(M_m^T Z_m^T Z_m M_{u_m} + tr(Z_m^T Z_m S_m)\right) \\ & + \frac{1}{2\sigma_\epsilon^2} E(B_m^T) Z_m M_m\Big\},\end{aligned} \tag{12}$$

where $w_m = E(q(r_m))$; $logit(w_m) = \log\left(\frac{w_m}{1-w_m}\right)$.

Update σ_m^2: The variational distribution for $q(\sigma_m^2)$ is $q(\sigma_m^2) = IG(a_m, b_m)$. The parameters are updated by

$$\begin{aligned} a_m &= \frac{n}{2} + a_1 \\ b_m &= E\left(\frac{1}{2} U_m^T U_m\right) + b_1\end{aligned} \tag{13}$$

Update σ_ϵ^2: The variational distribution for $q(\sigma_\epsilon^2)$ is $q(\sigma_\epsilon^2) = IG(a_\epsilon, b_\epsilon)$. The parameters a_ϵ and b_ϵ are updated according to following equations:

$$\begin{aligned} a_\epsilon &= \frac{n}{2} + a_0 \\ b_\epsilon &= E(\frac{1}{2}E(C^T C)) + b_0,\end{aligned} \tag{14}$$

where $C = y - X\Gamma\beta - \sum_{m=1}^{M}(Z_m(r_m I_m U_m))$.

3 Simulations

Simulation studies were conducted to evaluate the impact of disease models and the number of noise regions on the performance of the proposed method. For all simulation studies, to mimic the minor allele frequencies and linkage disequilibrium (LD) in the real genome, all genotype data was drawn directly from phase 1 (ADNI-1) and subsequent extensions (ADNI-GO/2) of ADNI study (i.e. $n = 808$) (Saykin *et al.*, 2015). We first excluded the genetic variants with more than 20% missing and then grouped them according to the gene annotation in GRch37 assembly. The average number of genetic variants for each gene is around 597 (range 10–4056), and the average length of the gene is 32.9 kb (range 2.6–303.1 kb). We further added a window of 5 kb upstream and downstream to each gene. In all our simulation studies, we treated each gene as a genetic region. To avoid over-fitting, 20% samples were randomly selected

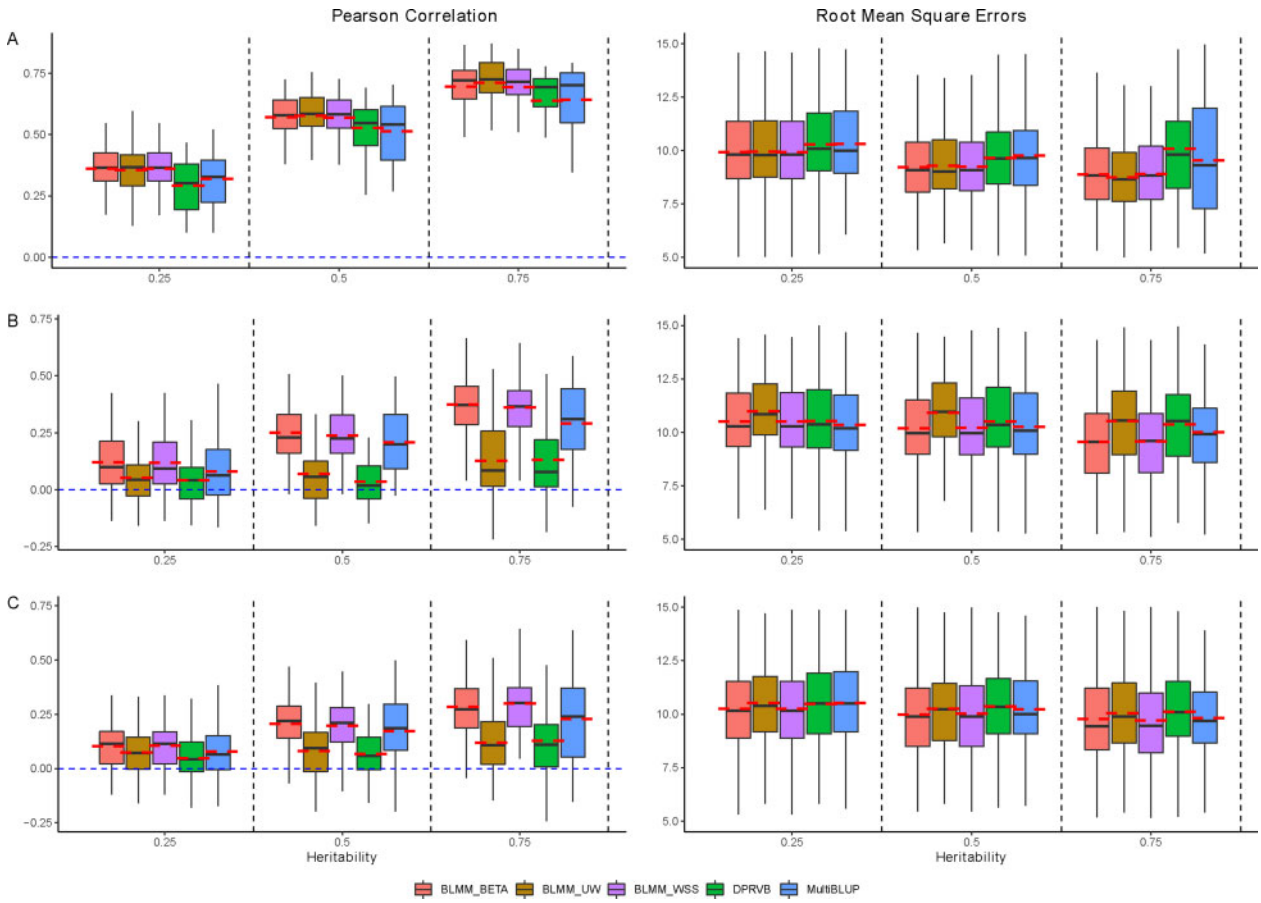

Fig. 1. The comparison of prediction accuracy when outcomes were caused by groups of predictors located nearby. (**A**) The outcomes were simulated with an equal weight matrix. (**B**) The outcomes were simulated with a BETA weight matrix. (**C**) The outcomes were simulated with a WSS weight matrix

for testing and the remaining was used for training. To reduce the chance finding problem, each simulation setting was replicated 100 times. For all methods, the prediction performance was evaluated by using the Pearson correlations and root mean square errors (RMSE).

We further compared our method with two widely used methods including MultiBLUP (Speed and Balding, 2014) and DPR with VB method (denoted as DPRVB). For computational reasons, we did not compare the performance of the proposed method to DPR with parameter estimated using the MCMC technique, which has similar performance as DPRVB but much slower convergence rate (Zeng and Zhou, 2017).

3.1 Scenario 1: the impact of disease model

In this set of simulations, we evaluated the performance of the proposed method under different disease models, including the outcomes were affected by (i) groups of predictors located nearby and (ii) isolated predictors.

3.1.1 The outcome is affected by predictors located nearby

We first evaluated the proposed method when diseases were caused by a number of predictors located in nearby regions (i.e. the assumption used in MultiBLUP). We randomly selected 30 genes from ADNI dataset and set 2 of them to be causal. We simulated the outcomes as

$$\boldsymbol{Y} \sim N(0, \sum_{i}^{R} \boldsymbol{K}_i \sigma_i^2 + \boldsymbol{I}\sigma_\epsilon^2), \tag{15}$$

where $\boldsymbol{K}_i = (\boldsymbol{G}_i \boldsymbol{W}_i \boldsymbol{G}_i^T)/p_i$; $\boldsymbol{G}_i$ is an $n \times p_i$ matrix of all genetic markers on gene i; and p_i is the number of variants on gene i. $\boldsymbol{W}_i$ is the weights that reflect the predictive effect of each genetic variant on the ith gene. We first considered the case where outcomes were mainly caused by common variants, and set $w_j = 1$ for each predictors. We further considered the scenario where rare variants contributed substantially to disease risk. We simulated two models under such settings, where a beta-type of weights $w_j = dbeta(MAF_j, 1, 25)^2$ and a weighted sum statistics types of weights $w_j = \frac{1}{MAF_j(1-MAF_j)}$ (denoted as WSS) were used. While the BLMM uses WSS weight to construct the variance-covariance structure for the random effects by default, other weights can also be employed. To evaluate the robustness of the default weight, we further analysed the simulated data using BLMM-UW and BLMM-BETA, where the un-weighted weights (i.e. $w_j = 1$) and beta types of weights were adopted respectively. For all the scenarios, we gradually changed the heritability from 25% to 75%.

The Pearson correlations and the RMSEs are shown inFigure 1. Not surprisingly, as the heritability increases, all methods tend to perform better. When the outcomes were simulated under the assumption that all genetic variants contributed equally (Fig. 1A), all BLMM methods tend to perform similarly. When the disease outcomes were simulated under the assumption that rare variants had large contributions (Fig. 1B and C), BLMM with BETA and WSS weights perform much better than the BLMM-UW. This is mainly because the weights in both the BLMM-BETA and BLMM-WSS are designed to capture the effects from rare variants. Indeed, the performance of BLMM-BETA and BLMM-WSS are very similar. Although BLMM with weights that reflect the underlying disease model performed the best, the BLMM-WSS performed close to the best under all the situations considered. As shown in Figure 1, BLMMs have much better performance than both MultiBLUP and DPRVB. This is mainly due to the fact that BLMM explicitly

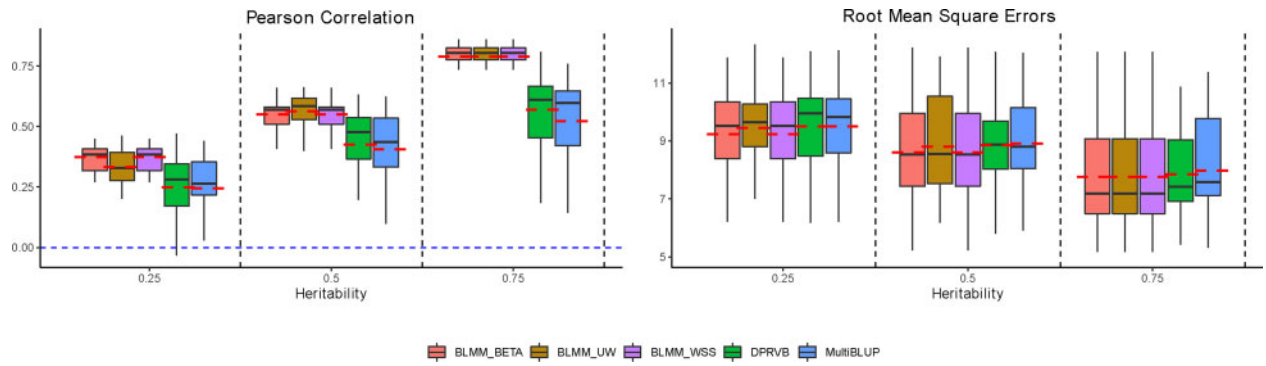

Fig. 2. The comparison of prediction accuracy when outcomes were affected by isolated predictors

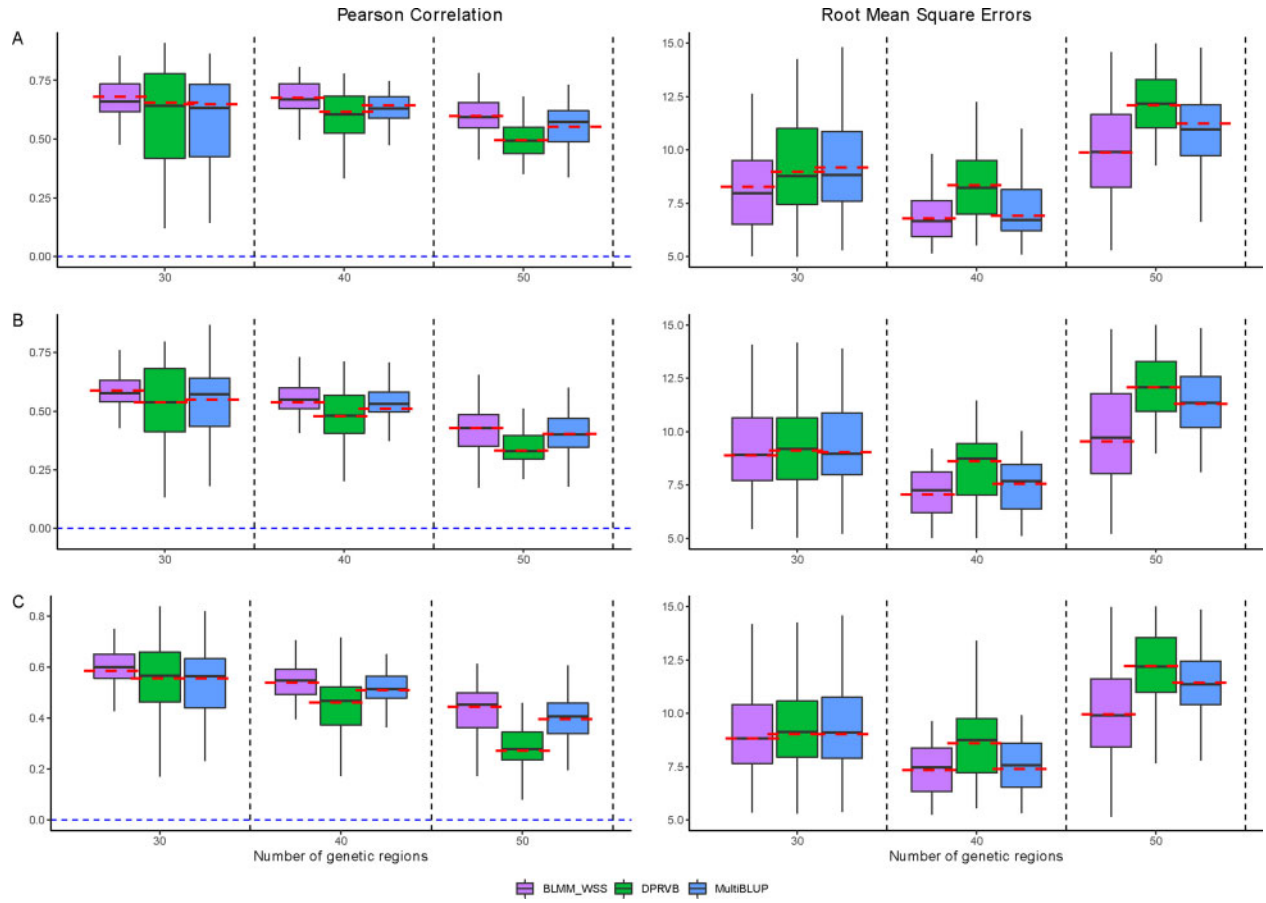

Fig. 3. The comparison of prediction accuracy under different number of noise regions. (**A**) The outcomes were simulated with an equal weight matrix and isolated genetic variants. (**B**) The outcomes were simulated with a BETA weight matrix and isolated genetic variants. (**C**) The outcomes were simulated with a WSS weight matrix and isolated genetic variants

accounts for the effects from rare variants and selects predictive regions, and thus substantially reduce the impact of noise. In addition, we noticed that DPRVB has much worse performance than both MultiBLUP and BLMM methods under current settings, where a group of predictors from nearby regions are predictive. This is partially because DPRVB assumes that only isolated markers have large predictive effects.

With regards to the variable selection, the averaged sensitivity and specificity for BLMM-WSS are 87% and 94%, respectively. The sensitivity and specificity for the other two weights (i.e. BLMM-UW and BLMM-BETA) are listed in Supplementary Table S1. Not surprisingly, BLMM with the weight that represents the underlying disease model has the best selection performance. However, the BLMM-WSS tends to be robust under various disease models. While BLMM-WSS performs the best when the model assumption was completely satisfied, its selection accuracy is close to the best that is obtained from BLMM with weights reflecting the true disease model (Supplementary Table S1).

3.1.2 The outcome is affected by isolated predictors

We further considered the case where the outcomes were caused by a small fraction of isolated genetic variants, the same assumption employed by the sparsity regression model. Similar to the above settings, 30 genes were drawn from the ADNI dataset. We randomly selected 1% of common genetic variants to serve as causal variants

and simulated the outcome as $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is a matrix of causal variants and $\boldsymbol{\beta} \sim \mathcal{N}(0, I\sigma_\beta^2)$ is their effects. Similar to Scenario I, we gradually increased the heritability from 25% to 75%.

The results are summarized in Figure 2. As the heritability increases, the performance of all methods improves. As expected, BLMM-WSS has the highest Pearson correlations and the smallest RMSEs across all the settings considered, and it captures most of the heritability. This is mainly due to the fact that our proposed method can not only model the effects from individual predictors but also select those that are predictive. DPRVB has slightly better performance than MultiBLUP, indicating the improved performance of DPRVB yields when its assumption is met.

With regards to the variable selection, the LD between causal and non-causal variants can have a big impact on the evaluation of selection accuracy (Berger *et al.*, 2015; Walters *et al.*, 2012; Yang *et al.*, 2010). As LD is mostly present within the gene and becomes negligible at distant locations (e.g. Supplementary Fig. S1), we calculated the probability of correctly select genes that harbour causal variants. BLMM-WSS can achieve an average of 80% sensitivity and 93% specificity (Supplementary Table S2). All genes have a relatively small chance of being selected from the random effects part of the BLMM model (Supplementary Table S2). This is consistent with the disease model, where the predictors are not located nearby.

3.2 Scenario 2: the impact of noise predictors

In this set of simulations, we evaluated the impact of the number of noise regions on the performance of the proposed method. We considered a disease model where diseases were caused by both isolated markers and a group of predictors located in nearby regions. Similar to Section 3.1, 2 genes were randomly selected to serve as groups of predictors located nearby, and 1% of common variants were set as causal isolated predictors. Regions without any causal markers served as the noise, and we gradually increased its number from 30 to 50. The outcomes were simulated as

$$\boldsymbol{Y} \sim N(\boldsymbol{\beta} + \sum_{i}^{R} \boldsymbol{g}_i, I\sigma_\epsilon^2) \tag{16}$$

where $\boldsymbol{g}_i \sim N(0, \boldsymbol{K}_i\sigma_i^2)$. The total heritability was set to be 80% with cumulative genetic effects and isolated additive effects each contributing half of the genetic variance (i.e. $Var(\boldsymbol{\beta}) = Var(\sum_R \boldsymbol{g}_i)$). Same as Section 3.1.1, for cumulative effects, we used three weight matrices to allow rare variants having various levels of contributions to disease risks.

The Pearson correlations and RMSEs are shown in Figure 3, and the computational time (Intel® Xeon® Processor E5-2695 v4 2.1 GHz, dual core) and memory as the number of noise regions increases is shown in Supplementary Figure S2. Regardless of the number of regions and the underlying disease models, the proposed method performs better than the others. As the number of noise regions increases, the prediction accuracy from all methods decreases. However, the performance of the other methods dropped much faster as compared to the BLMM method. This is partially due to the fact that the proposed method is capable of selecting and modelling the predictive effects from both individual variants and genes, which substantially reduces the impact of noise regions/predictors. MultiBLUP uses empirical criteria to select predictive regions, and thus has improved performance as compared to DPRVB. However, it can't capture the effects from isolate predictors and the adopted empirical criteria may reduce the robustness of its performance. For DPRVB, it uses truncated stick-breaking approximation approach to select the normal components for the density estimation, which is not directly related to the selection of predictive variables. Their strategy incurs uncertainty in selecting predictors which can negatively affect models' predictive ability.

With regards to the variable selection, the BLMM-WSS can reasonably detect both causal variants and regions. As the number of noise region increases from 30 to 50, the average sensitivity changed from 74% to 63% and the specificity changed from 92% to 85%. The detailed selection performance for additive and random effects for the BLMM model under each disease model is shown in Supplementary Table S3.

4 Real data application

We were interested in predicting positron emission tomography (PET) imaging outcomes using the whole-genome sequencing data from the ADNI study. ADNI is a multi-site study designed for the prevention and treatment of Alzheimer's Disease (AD) (Mueller *et al.*, 2005). The initial phase of ADNI (ADNI-1) and its subsequent phases (ADNI-GO/2) recruit participants aged from 55 to 90, at 57 sites across the United States and Canada (Nho *et al.*, 2016). The whole-genome sequencing (WGS) was performed on blood-derived genomic DNA samples from 818 ADNI (1/GO/2) participants (Saykin *et al.*, 2015). Samples were sent to a non-Clinical Laboratory Improvements Amendments (non-CLIA) laboratory at Illumina and sequenced on the Illumina HiSeq2000 (Saykin *et al.*, 2015). After discarding the related individuals, 808 subjects remained for my analysis (Petersen *et al.*, 2010). To ensure data quality, variants not successfully genotyped (missing rate > 1%) were filtered out from the raw data.

The distribution of the PET-imaging outcomes, including florbetapir (AV45, Mean = 1.21, STD = 0.23) and fluorodeoxy glucose (FDG, Mean = 6.04, STD = 0.8), are shown in Supplementary Figure S3. The sample sizes for FDG and AV45 are 639 and 501, respectively. We used 57 susceptibility genes (Supplementary Table S4) that have been reported to be associated with AD, and included a total of 115 749 genetic variants in our analysis. To avoid overfitting, 80% of the data was used to train the model. Pearson correlations and RMSEs were calculated based on the remaining 20% of the data. To avoid the chance findings, this process was replicated 100 times. For comparison purposes, we also built prediction models using the DPRVB and MultiBLUP methods.

The prediction accuracies are shown in Figure 4. BLMM-WSS outperformed both the MultiBLUP and DPRVB methods for the prediction of FDG and has comparable performance to that of MultiBLUP for the prediction of AV45. For both outcomes, DPRVB

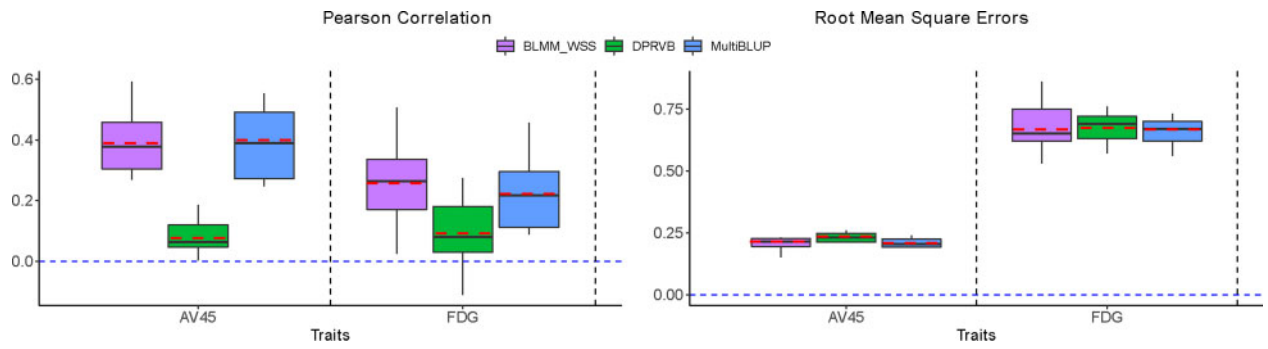

Fig. 4. The prediction accuracies for positron emission tomography imaging outcomes

that is not capable of selecting predictors performs substantially worse than both BLMM and MultiBLUP, indicating variable selections can be of great importance for an accurate and robust prediction model. To further explore our model, the probability of each gene being selected was summarized (Supplementary Table S4). Three genes (i.e. *APOC, APOE and TOMM40*) have been selected frequently for both AV45 and FDG. Many previous studies have shown that these three genes play an important role in AD. For examples, Ossenkoppele *et al.* (2013) found that *APOE ε4* allele has a differential effect on glucose metabolism in AD patients. APOs (e.g. *APOC-III* and *APOE*) have also been confirmed to be associated with AD-related pathologies (Zou *et al.*, 2019). *TOMM40* and *APOE* are related to the late-onset AD (Roses, 2010).

5 Discussion and conclusions

We have presented a novel BLMM framework for risk prediction analysis using high-dimensional genetic data. The proposed framework is (i) flexible and robust against various disease models (i.e. diseases can be affected by predictors located nearby and/or isolated); (ii) capable of selecting and capturing the predictive effects from both common and rare variants; and (iii) less sensitive to the number of noise predictors. Through extensive simulation studies, we have demonstrated that BLMM has higher prediction accuracy as compared to the commonly used MultiBLUP and DPRVB (Speed and Balding, 2014; Zeng and Zhou, 2017) and can correctly select predictive variants/regions. Moreover, although BLMM performs the best when weights reflect the underlying disease model, the default weight (i.e. BLMM-WSS) has the best or close-to-the-best performance. We considered this important, as in practice the disease model is usually unknown in advance.

Many studies have demonstrated that rare variants can improve the accuracies of risk prediction models (Gibson, 2012; Taudien *et al.*, 2016). However, their large number, low minor allele frequencies and unknown types of effects pose significant challenges for prediction modelling. Most existing methods are designed for common variants (Speed and Balding, 2014; Weissbrod *et al.*, 2016; Zeng and Zhou, 2017; Zhou *et al.*, 2013), and thus are not capable of capturing the predictive effects from rare variants. In this study, we used a similar idea in the association analysis of sequencing data (Speed and Balding, 2014; Wu *et al.*, 2011), and assumed that genetic similarity leads to phenotypic similarity. We incorporated a WSS weight into the genetic similarity measure, making it capable of capturing the predictive effects from rare variants. As shown in simulation studies (Fig. 1), this strategy yields improved prediction accuracy when rare variants contribute to disease risk. While the ideal choice of weights should reflect the underlying disease model, the proposed WSS weight has the best or close-to-optimal performance across a wide range of settings (Fig. 1).

Most of the existing methods either assume that diseases are caused by a few isolate predictors(Carvalho *et al.*, 2008; Zhou *et al.*, 2013) or a group of predictors located nearby(Speed and Balding, 2014; Weissbrod *et al.*, 2016; Zeng and Zhou, 2017) and thus their performance can be affected by the unknown underlying disease model. The proposed BLMM can be viewed as a unified framework, where both of these commonly used assumptions are accommodated. Specifically, the random effect part of BLMM is designed to capture the cumulative effects from predictors including rare variants in nearby regions, whereas the fixed effect part is for modelling the effects from isolated predictors. As shown in simulations, the proposed method achieves the highest prediction accuracy among all the methods considered regardless of the underlying disease models.

For high-dimensional sequencing data, a large amount of measured genetic variants are not predictive, and thus variable selection is of great importance for building accurate prediction models. However, the majority of the existing variable selection methods are designed for selecting isolated predictors (Habier *et al.*, 2011; Yi and Xu, 2008; Zeng and Zhou, 2017; Zhou *et al.*, 2013) and few can be used for selecting groups of predictions. While the LMM-based prediction models have the capacity for selecting groups of predictors located nearby (Speed and Balding, 2014; Weissbrod *et al.*, 2016), they usually use empirical criteria that can lead to suboptimal prediction performance. Contrary to existing methods, the proposed BLMM uses 'Bernoulli-Gaussian prior' and 'spike and slab prior' to facilitate the variable selection. It can not only select predictive markers but also regions with many predictors harboured. It achieves over 80% sensitivity and specificity across all the simulation settings (Supplementary Tables S1–S3). With noise regions/variants being excluded, the BLMM has a much better prediction accuracy as compared to MultiBLUP and DPRVB.

Through the analysis of PET imaging outcomes, we found that the overall accuracy of BLMM-WSS is greater compared to the other methods. *APOE*, *APOC* and *TOMM40* located on chromosome 19 are frequently selected. Mounting evidences have indicated that these three genes are important causative elements of AD. Polymorphism of *APOE* can influence the neuronal repair mechanisms and the maintenance of synaptic connections (Ferencz *et al.*, 2012). *TOMM40* is considered as a promising lending gene in AD onset and plays an essential role in mitochondrial survival (Ferencz *et al.*, 2012). Recent reports show that *APOE* locus have LD patterns with *APOE* on 19q13.2 (Bekris *et al.*, 2012). This strong LD suggests it is difficult to establish the association between risk of AD and *APOE* promoter polymorphisms independent (Bekris *et al.*, 2008). SNPs at the *TOMM40* gene have been reported to be associated with higher cerebrospinal fluid and post-mortem brain apolipoprotein E (*apoE*) expression in the hippocampus of AD patients (Bekris *et al.*, 2012). *APOC* polymorphism has been reported to be associated with an increased risk of the late-onset AD, and it also has interaction with *APOE* (Martins *et al.*, 2009).

Similar to existing literature, the proposed BLMM only applies to continuous outcomes. While conventional strategies normally treat binary outcomes as if they were continuous (Speed and Balding, 2014; Weissbrod *et al.*, 2016; Zeng and Zhou, 2017; Zhou *et al.*, 2013), it would be of interest to develop a generalized Bayesian LMM framework that can explicitly take the distribution of the outcome into consideration. Currently, BLMM only focuses on additive effects, and it would be natural to extend BLMM for non-additive effects (e.g. dominance and epistasis effects), where various kernel functions (e.g. quadratic and radial basis kernels) can be used. Although KPCA and VB algorithms were used to reduce the computation cost, it can still be computationally expensive for genome-wide data, especially when various types of effects (i.e. additive or pairwise interaction) are considered. These will be the future directions of my research.

In summary, we have developed a novel BLMM method. It can capture the predictive effects from both common and rare variants, and provide robust prediction performance against various underlying disease models. We considered the proposed BLMM model as the efficient apparatus for use in a wide range of risk prediction tasks.

Acknowledgements

The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and/or training services as part of this research.

Funding

This project was funded by the Faculty Research Development Fund from the University of Auckland, the Marsden Fund from Royal Society of New Zealand (Project No. 19-UOA-209) and the National Library of Medicine (Award No. 1R01LM012848-01). New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme (https://www.nesi.org.nz).

Conflict of Interest

none declared.

References

Algamal,Z.Y. and Lee,M.H. (2015) Penalized logistic regression with the adaptive lasso for gene selection in high-dimensional cancer classification. *Expert Syst. Appl.*, **42**, 9326–9332.

Ashley,E.A. (2015) The precision medicine initiative: a new national effort. *J. Am. Med. Assoc.*, **313**, 2119–2120.

Bekris,L.M. *et al.* (2008) Multiple SNPs within and surrounding the apolipoprotein E gene influence cerebrospinal fluid apolipoprotein e protein levels. *J. Alzheimers Dis.*, **13**, 255–266.

Bekris,L.M. *et al.* (2012) Functional analysis of APOE locus genetic variation implicates regional enhancers in the regulation of both TOMM40 and APOE. *J. Hum. Genet.*, **57**, 18–25.

Berger,S. *et al.* (2015) Effectiveness of shrinkage and variable selection methods for the prediction of complex human traits using data from distantly related individuals. *Ann. Hum. Genet.*, **79**, 122–135.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, Singapore.

Blei,D.M. *et al.* (2017) Variational inference: a review for statisticians. *J. Am. Stat. Assoc.*, **112**, 859–877.

Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.

Byrnes,A.E. *et al.* (2013) The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet. Epidemiol.*, **37**, 666–674.

Carbonetto,P. *et al.* (2012) Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, **7**, 73–108.

Carvalho,C.M. *et al.* (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438–1456.

Chatterjee,N. *et al.* (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, **17**, 392–406.

Chen,Z. and Dunson,D.B. (2003) Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769.

De los Campos,G. *et al.* (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.*, **9**, e1003608.

Dunson,D.B. (2001) Commentary: practical advantages of Bayesian analysis of epidemiologic data. *Am. J. Epidemiol.*, **153**, 1222–1226.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Ferencz,B. *et al.* (2012) Promising genetic biomarkers of preclinical alzheimer's disease: the influence of APOE and TOMM40 on brain integrity. *Int. J. Alzheimer's Dis.*, **2012**, 1–15.

Fernandes,V. *et al.* (2017) Bernoulli–Gaussian distribution with memory as a model for power line communication noise. In *Proc. Braz. Telecommun. Signal Process. Symp.*, São Pedro, pp. 328–332.

Ghahramani,Z. and Beal,M.J. (2000) Variational inference for Bayesian mixtures of factor analysers. In: *Advances in Neural Information Processing Systems*, November 29–December 4, 1999, Denver, CO, USA, pp. 449–455.

Gianola,D. (2013) Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*, **194**, 573–596.

Gibson,G. (2012) Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, **13**, 135–145.

Habier,D. *et al.* (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**, 186.

Hayes,B. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

,Henderson, C.R. (1950) Estimation of Genetic Parameters. *Ann. Math. Stat.*, **21**, 309–310.

Henderson,C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

Hoggart,C.J. *et al.* (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Lipinski,K.A. *et al.* (2016) Cancer evolution and the limits of predictability in precision cancer medicine. *Trends Cancer*, **2**, 49–63.

Martins,I.J. *et al.* (2009) Cholesterol metabolism and transport in the pathogenesis of Alzheimer's disease. *J. Neurochem.*, **111**, 1275–1308.

Mitchell,T.J. and Beauchamp,J.J. (1988) Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, **83**, 1023–1032.

Moser,G. *et al.* (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Select. Evol.*, **41**, 56.

Mueller,S.G. *et al.* (2005) The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin.*, **15**, 869–877.

Nho,K. *et al.*; ADNI. (2016) Integration of bioinformatics and imaging informatics for identifying rare PSEN1 variants in Alzheimer's disease. *BMC Med. Genomics*, **9**, 30.

Ormerod,J.T. *et al.* (2017) A variational Bayes approach to variable selection. *Electronic J. Stat.*, **11**, 3549–3594.

Ossenkoppele,R. *et al.* (2013) Differential effect of APOE genotype on amyloid load and glucose metabolism in ad dementia. *Neurology*, **80**, 359–365.

Petersen,R. *et al.* (2010) Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. *Neurology*, **74**, 201–209.

Roses,A.D. (2010) An inherited variable poly-t repeat genotype in tomm40 in Alzheimer disease. *Arch. Neurol.*, **67**, 536–541.

Saint Pierre,A. and Genin,E. (2014) How important are rare variants in common disease? *Brief. Funct. Genomics*, **13**, 353–361.

Salimans,T. *et al.* (2015) Markov chain Monte Carlo and variational inference: Bridging the gap. In: *International Conference on Machine Learning*, Lille, France, 07–09 Jul 2015, 37, pp. 1218–1226.

Saykin,A.J. *et al.* (2015) Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans. *Alzheimer's & Dementia*, **11**(7), 792–814.

Schork,A.J. *et al.*; The Tobacco and Genetics Consortium. (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.*, **9**, e1003449.

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

Taudien,S. *et al.* (2016) Genetic factors of the disease course after sepsis: rare deleterious variants are predictive. *EBioMedicine*, **12**, 227–238.

Cerquides,J. and de Mantaras,R.L. (2003) The indifferent naive Bayes classifier.*Proceedings of the 16th International FLAIRS Conference*, May 12–14, 2003, St. Augustine, FL, USA, pp. 341–345.

Walters,R. *et al.* (2012) An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data. *Bioinformatics*, **28**, 2615–2623.

Weissbrod,O. *et al.* (2016) Multikernel linear mixed models for complex phenotype prediction. *Genome Res.*, **26**, 969–979.

Weissfeld,J.L. *et al.* (2015) Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. *J. Thoracic Oncol.*, **10**, 1538–1545.

Wen,Y. *et al.* (2016) Risk prediction modeling of sequencing data using a forward random field method. *Sci. Rep.*, **6**, 21120.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

Yi,N. and Xu,S. (2008) Bayesian lasso for quantitative trait loci mapping. *Genetics*, **179**, 1045–1055.

Zeng,P. and Zhou,X. (2017) Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.*, **8**, 456.

Zhang,C. *et al.* (2019) Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, **41**, 2008–2026.

Zhao,Y. *et al.* (2006) General design Bayesian generalized linear mixed models. *Stat. Sci.*, **21**, 35–51.

Zhou,X. *et al.* (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.*, **9**, e1003264.

Zou,S. *et al.* (2019) Subtypes based on six apolipoproteins in non-demented elderly are associated with cognitive decline and subsequent tau accumulation in cerebrospinal fluid. *J. Alzheimer's Dis.*, **72**, 413–423.