frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 19 October 2023
DOI 10.3389/fgene.2023.1267704

OPEN ACCESS

EDITED BY
Tao Wang,
Medical College of Wisconsin,
United States

REVIEWED BY
Song Zhai,
Merck, United States
Zhaoxia Yu,
University of California, Irvine,
United States

*CORRESPONDENCE
Long Liu,
biostat-ll@sxmu.edu.cn
Yalu Wen,
y.wen@auckalnd.ac.nz

†These authors have contributed equally to this work

RECEIVED 26 July 2023
ACCEPTED 25 September 2023
PUBLISHED 19 October 2023

CITATION
Hai Y, Zhao W, Meng Q, Liu L and Wen Y (2023), Bayesian linear mixed model with multiple random effects for family-based genetic studies.
Front. Genet. 14:1267704.
doi: 10.3389/fgene.2023.1267704

Bayesian linear mixed model with multiple random effects for family-based genetic studies

Yang Hai[1†], Wenxuan Zhao[2†], Qingyu Meng[2], Long Liu[2*] and Yalu Wen[1*]

[1]Department of Statistics, University of Auckland, Auckland, New Zealand, [2]Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, China

Motivation: Family-based study design is one of the popular designs used in genetic research, and the whole-genome sequencing data obtained from family-based studies offer many unique features for risk prediction studies. They can not only provide a more comprehensive view of many complex diseases, but also utilize information in the design to further improve the prediction accuracy. While promising, existing analytical methods often ignore the information embedded in the study design and overlook the predictive effects of rare variants, leading to a prediction model with sub-optimal performance.

Results: We proposed a Bayesian linear mixed model for the prediction analysis of sequencing data obtained from family-based studies. Our method can not only capture predictive effects from both common and rare variants, but also easily accommodate various disease model assumptions. It uses information embedded in the study design to form surrogates, where the predictive effects from unmeasured/unknown genetic and environmental risk factors can be modelled. Through extensive simulation studies and the analysis of sequencing data obtained from the Michigan State University Twin Registry study, we have demonstrated that the proposed method outperforms commonly adopted techniques.

Availability: R package is available at https://github.com/yhai943/FBLMM.

KEYWORDS

bayesian linear mixed model, family-based genetic study, rare variants, unknown genetic factors, common environmental risk factors

Introduction

Family-based study (e.g., twin study) is one of the most popular designs used in genetic research, and it offers many unique features for risk prediction studies. For example, the relatedness among family members helps capture the predictive effects from unmeasured/unknown polygenic and shared environmental factors, and thus contributes additional information, beyond the measured data, for risk prediction studies (Ruderfer et al., 2010). Despite these advantages, few statistical methods are available for risk prediction research using family-based designs. The existing methods usually build risk prediction models based on genetic effects that are estimated with familial correlations adjusted for. For example (Meigs et al., 2008), developed a risk prediction model for family-based genetic studies, where the genotypic risk score is determined without considering the information in families (Ruderfer et al., 2010). presented a family-based liability threshold model and illustrated it in

the analyses of Crohn's disease. Although these methods have contributed to the advances of family-based risk prediction, they can lead to less accurate models when unmeasured genetic and/or shared environmental factors contribute significantly to disease risk. Moreover, the recent whole genome sequencing studies have demonstrated that rare variants can play a significant role in many common complex diseases, such as obesity, coronary heart disease, and drug addiction (Ramachandrappa et al., 2013; Peloso et al., 2014; Wang et al., 2014). Family-based design can enhance the chance of capturing the predictive effects from rare variants as they tend to be aggregated within family. However, existing prediction models do not utilize the design information and they simply extend models designed for population-based studies by adjusting correlations within the data. Therefore, it remains challenging for them to capture the predictive effects from rare variants, primarily due to their low minor allele frequencies (Mihaescu et al., 2013).

It has long been recognized that family history alone can greatly facilitate disease risk prediction. For many complex diseases (e.g., cardiovascular diseases and type II diabetes), individuals with a positive family history are usually classified as the population at high risk (Valdez et al., 2007; Marateb et al., 2018). Family history can be viewed as a surrogate that reflects the contributions of many known/unknown risk factors accumulated within a family. Evidences have shown that familial effects account for a significant amount of disease variability. For example (Chen et al., 2007), have shown that 33% of variance of spherical equivalent can be attributed to childhood environmental effects. Furthermore, genetic variants can account for a substantial proportion of heritability for human traits (Couillard et al., 2001; Dirani et al., 2006). For example, genetic factors can explain as much as 87% of the variation in the susceptibility to asthma in twins with positive family history (Laitinen et al., 1998; Lichtenstein et al., 2009) found that genetic heritability for bipolar disorder and schizophrenia was 59% and 64%, respectively. The familial aggregation for many complex diseases is mainly due to the relatedness in genetic and environmental factors among family members, which carry important information and can be used to further improve prediction accuracy. However, most existing analytical methods are developed by simply extending those models designed for population-based studies, where family correlations are first adjusted. For example (Meigs et al., 2008), built a risk prediction model for family-based genetic study, where the relatedness among family members is adjusted using a generalized estimating equation model. Although statistically valid and these methods could capture the predictive effects from those measured known risk factors, they are not capable of using family information as surrogates to account for unmeasured predictors (e.g., shared environmental risk factors).

Population-based whole-genome sequencing studies have shown that rare variants are associated with many complex human diseases (Dickson et al., 2010; Helgadottir et al., 2016), and they have great potential in explaining the missing heritability (Cruceanu et al., 2013). For example, recent study has reported that rare variants in renal salt handling genes have contributed to variation of blood pressure (Ji et al., 2008; Stefansson et al., 2008) found that rare variants are associated with schizophrenia and autism (Ionita-Laza and Ottman, 2011). showed that four rare variants in *IFIH1* gene can lower the risk of type 1 diabetes. Recent developments in prediction research have also shed light on the importance of rare variants in building an accurate prediction model. For example, the risk prediction model for coronary artery disease in European and South Asian populations was built with rare variants incorporated, and it yields improved predictive accuracy (Lali et al., 2020). Despite their importance, few methods designed for family-based studies have considered the contributions of rare variants in disease risk prediction. Recently, we developed a Bayesian linear mixed model with multiple random effects (denoted as BLMM) to predict disease risk for population-based studies, where both common and rare variants have been explicitly considered (Hai and Wen, 2020). We have showed that the BLMM can capture the predictive effects from rare variants and is robust against various disease models. Though promising, it was developed for population-based studies, and thus cannot make use of the information embedded in the family-based study design.

To address these limitations, we proposed a family-based Bayesian linear mixed model with multiple random effects (denoted as FBLMM) for the prediction analysis on sequencing data obtained from family-based genetic studies. The proposed FBLMM uses the correlations among family members to construct surrogates for unmeasured risk predictors, and it can account for the predictive effects from both common and rare variants. In the following sections, we first presented the details of the proposed model, and then conducted extensive simulation studies to evaluate its performance. Finally, we illustrated its application using the whole-genome exome data from Michigan State University Twin Registry study (Burt and Klump, 2012).

Materials and methods

The proposed FBLMM is built using a similar idea in BLMM presented in (Hai and Wen, 2020), where we assume genetic similarities can lead to phenotypic similarities. Fundamentally different from existing methods that adjust for family correlations, we utilize the information embedded in the family-based study design to further improve the prediction accuracy. Given M genetic regions that can be defined using various criteria (e.g., gene and pathway), we form the FBLMM model as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{m=1}^{M} \mathbf{g}_m + \mathbf{f} + \boldsymbol{\epsilon}_n \quad \text{with} \quad \boldsymbol{\epsilon}_n \sim \mathcal{N}\left(0, \mathbf{I}\sigma_\epsilon^2\right), \tag{1}$$

where $\mathbf{Y}$ is the outcome; $\mathbf{X}$ is the genotypes for all common variants; and $\boldsymbol{\beta}$ is their corresponding effect. $\mathbf{g}_m$ is the cumulative predictive effect from all measured predictors, including rare variants, on region m. $\mathbf{f}$ is the familial effect due to shared environmental factors and genetic relatedness, and $\mathbf{I}$ is an $n \times n$ identity matrix.

Similar to existing sparsity regression models (Carvalho et al., 2008; Zhou, Carbonetto, and Stephens, 2013), the $\mathbf{X}\boldsymbol{\beta}$ is designed to capture the predictive effects from isolated markers. To tease out the impact of noise, we followed the same procedure in (Hai and Wen, 2020), instead of using the spike and slab prior that can lead to an underestimation of posterior variances for $\boldsymbol{\beta}$ (Carbonetto et al., 2012). We re-parameterized $\mathbf{X}\boldsymbol{\beta}$ as $\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}$, where $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ is a vector of binary variables indicating whether each genetic variant is predictive. We used the Bernoulli Gaussian distribution as the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (i.e., $\beta_j \sim \mathcal{N}(0, \sigma_\beta^2)$ and $\gamma_j \sim \text{Bernoulli}(\theta_0 = 0.1)$), and this allows to obtain an unbiased estimation of the posterior variance of $\boldsymbol{\beta}$ as well as achieving variable

selection for $\boldsymbol{\beta}$ (Zhou, Carbonetto, and Stephens, 2013; Fernandes et al., 2017).

Similar to linear mixed models that assume the infinitesimal effects (VanRaden, 2008), the cumulative predictive effects from common and rare variants for region m are modeled via $\mathbf{g}_m$, where we set a multivariate normal prior for each region-based cumulative predictive effect as

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

$\mathbf{K}_m$ is the genetic similarity for region m and it is defined as $\mathbf{K}_m = \mathbf{G}_m\mathbf{W}_m\mathbf{G}_m^T/p_m$, where $\mathbf{G}_m$ is the genotype matrix for region m and p_m is the number of genetic markers in the region. $\mathbf{W}_m = (w_1, w_2, ..., w_{p_m})^T$ is the pre-specified weights used to capture the contribution of rare variants. Similar to existing literature (Wu et al., 2011; Lee et al., 2012), we define the weighted sum statistics types (denoted as WSS) of weights as $w_j = \frac{1}{MAF_j(1-MAF_j)}$, where MAF_j is the minor allele frequency for the j th variant. The hyper-parameters of a_1 and b_1 are set to be 0.1 for all regions. To expedite its computation, we re-parameterized the cumulative predictive effects part with the slab and spike prior as

$$\mathbf{Y} = \mathbf{X\Gamma\beta} + \sum_{m=1}^{M}(\mathbf{Z}_m r_m \mathbf{U}_m) + \mathbf{f} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{U}_m \sim N(0, \mathbf{I}\sigma_m^2)$, $\mathbf{Z}_m = \mathbf{Q}_m\boldsymbol{\Lambda}_m^{\frac{1}{2}}$ and $E(r_m = 1) = \wp(r_m)$. The re-parameterization facilitates the selection of predictive regions (i.e., $r_m = 1$ indicates the region is predictive), and the details of its derivations can be found in appendix A.

Mounting evidences suggest that there are familial aggregations for many complex traits (Laitinen et al., 1998; Couillard et al., 2001; Dirani et al., 2006; Lichtenstein et al., 2009), and the relatedness in genetic and environmental factors among family members are thought to be the main reasons for this aggregation. Therefore, we split the familiar effect $\mathbf{f}$ into predictive effects due to genetic correlation (denoted as $\mathbf{g}_f$) and shared environmental factors (denoted as $\mathbf{e}_f$). Model 3 can be written as

$$\mathbf{Y} = \mathbf{X\Gamma\beta} + \sum_{m=1}^{M}(\mathbf{Z}_m r_m \mathbf{U}_m) + \mathbf{g}_f + \mathbf{e}_f + \boldsymbol{\epsilon}_n. \tag{4}$$

We set the prior for $\mathbf{g}_f$ as

$$\begin{array}{rl} \mathbf{g}_f & \sim \mathcal{N}\left(0, \mathbf{K}^{g_f}\sigma_{g_f}^2\right), \\ \sigma_{g_f}^2 & \sim IG\left(a_{0f}, b_{0f}\right) \end{array} \tag{5}$$

where $\mathbf{K}^{g_f}$ is the theoretical kinship coefficient matrix. The $\mathbf{g}_f$ uses the genetic correlation between family members to improve the prediction accuracy, and it can be viewed as a surrogate for those predictive but unmeasured genetic variants. To account for the impact of environmental factors, we assume all family members share the same environment (e.g., diet) and set $\mathbf{e}_f$ as

$$\mathbf{e}_f \sim \mathcal{N}\left(0, \mathbf{K}^{e_f}\sigma_{e_f}^2\right), \tag{6}$$

where $\mathbf{K}^{e_f}$ is a block diagonal matrix with each block being a matrix with all elements equal to 1. The $\mathbf{e}_f$ is designed to capture the predictive effects from shared environmental factors, and it can also be viewed as a surrogate for those unmeasured environmental predictors shared by family members. We used the idea from (Z. Chen and Dunson, 2003) and decomposed $\mathbf{K}^{e_f}$ and $\mathbf{K}^{g_f}$ as $\mathbf{K}^{e_f} = \mathbf{Q}_{ef}\boldsymbol{\Lambda}_{ef}\mathbf{Q}_{ef}^T$ and $\mathbf{K}^{g_f} = \mathbf{Q}_{gf}\boldsymbol{\Lambda}_{gf}\mathbf{Q}_{gf}^T$, where $\boldsymbol{\Lambda}_{ef}$ and $\boldsymbol{\Lambda}_{gf}$ are diagonal matrices with eigenvalues on their diagonals, and $\mathbf{Q}_{ef}$ and $\mathbf{Q}_{gf}$ are matrices of the corresponding eigenvectors. Eq. 4 can be written as

$$\mathbf{Y} = \mathbf{X\Gamma\beta} + \sum_{m=1}^{M}(\mathbf{Z}_m r_m \mathbf{U}_m) + \mathbf{Z}_{gf}\mathbf{U}_{gf} + \mathbf{Z}_{ef}\mathbf{U}_{ef} + \boldsymbol{\epsilon}_n, \tag{7}$$

where $\mathbf{Z}_{gf} = \mathbf{Q}_{gf}\boldsymbol{\Lambda}_{gf}^{\frac{1}{2}}$ and $\mathbf{Z}_{ef} = \mathbf{Q}_{ef}\boldsymbol{\Lambda}_{ef}^{\frac{1}{2}}$. We adopted the mean-field variational Bayes algorithm (VB) to estimate parameters for FBLMM. Let $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{U}^g, \mathbf{U}_{gf}, \mathbf{U}_{ef}, \mathbf{r}, \boldsymbol{\sigma}^2, \sigma_{gf}^2, \sigma_{ef}^2, \sigma_{\boldsymbol{\epsilon}}^2)$ denotes all parameters of interest, where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)$, $\mathbf{U}^g = (\mathbf{U}_1, \cdots, \mathbf{U}_M)$, $\mathbf{r} = (r_1, \cdots, r_M)$, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_M^2)$. The goal is to obtain an optimal approximation $q(\boldsymbol{\xi})$ of the posterior distribution on $\boldsymbol{\xi}$ by maximizing the evidence lower bound (ELBO). In details, we iteratively update the approximated distributions for $q(\boldsymbol{\xi})$ as

$$\begin{aligned} q(\boldsymbol{\xi}) = \; & q_{\boldsymbol{\beta}} \times \prod_{j=1}^{p} q_{\gamma_j} \times \prod_{m=1}^{M} q_{\mathbf{U}_m} \times \prod_{m=1}^{M} q_{r_m} \times q_{\mathbf{U}_{gf}} \times q_{\mathbf{U}_{ef}} \\ & \times \prod_{m=1}^{M} q_{\sigma_m^2} \times q_{\sigma_{gf}^2} \times q_{\sigma_{ef}^2} \times q_{\sigma_{\boldsymbol{\epsilon}}^2}, \end{aligned} \tag{8}$$

where $q_{\boldsymbol{\beta}} = \mathcal{N}(\mathbf{M}_{\boldsymbol{\beta}}, \mathbf{S}_{\boldsymbol{\beta}})$; $q_{\gamma_j} = Bernoulli(\psi_j)$; $q_{\mathbf{U}_m} = \mathcal{N}(\mathbf{M}_m, \mathbf{S}_m)$; $q_{r_m} = Bernoulli(\phi_m)$; $q_{\mathbf{U}_{gf}} = \mathcal{N}(\mathbf{M}_{gf}, \mathbf{S}_{gf})$; $q_{\mathbf{U}_{ef}} = \mathcal{N}(\mathbf{M}_{ef}, \mathbf{S}_{ef})$; $q_{\sigma_m^2} = IG(a_m, b_m)$; $q_{\sigma_{gf}^2} = IG(a_{gf}, b_{gf})$; $q_{\sigma_{ef}^2} = IG(a_{ef}, b_{ef})$; and $q_{\sigma_{\boldsymbol{\epsilon}}^2} = IG(a_{\boldsymbol{\epsilon}}, b_{\boldsymbol{\epsilon}})$. Each parameter of $\boldsymbol{\xi}$ can be estimated by using the coordinate ascent algorithm, the estimating equations used to update the parameters are listed in appendix A.

The pseudo-code implementing our proposed model is shown in Figure 1. It is worth noting that when a new subject is not from families in the training data, its predicted value only depends on demographic and genetic predictors (i.e., the family information does not contribute to the outcomes). When a new individual comes from families in the training set, the FBLMM method not only uses genetic and demographic predictors, but also utilizes the extra information provided by family design to capture unmeasured genetic and shared environmental risk factors. Therefore, FBLMM has great potential to further improve predictions. The weight function employed by FBLMM can facilitate the identification of rare variants that are predictive, enabling FBLMM to consider contributions from both common and rare variants in prediction modeling.

Simulation study

We conducted extensive simulation studies to evaluate the performance of our proposed method under various family-based designs, and further compared FBLMM with other widely used methods, including 1) adaptive MultiBLUP (Speed and Balding, 2014); 2) DPRVB (Zeng and Zhou, 2017); and 3) BLMM (Hai and Wen, 2020), where family correlations are first adjusted. Note that both MultiBLUP and DPRVB have shown to outperform other existing gBLUP-based methods (Speed and Balding, 2014; Zeng and Zhou, 2017).

To closely mimic the real human genome, the founders' genotypes were drawn directly from Alzheimer's Disease

Algorithm 1: Inference procedure using variational Bayes.

Input: $\boldsymbol{X}, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_M, \boldsymbol{K}_{gf}, \boldsymbol{K}_{ef}, \boldsymbol{y}$

Output: $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{U}^g, \hat{\boldsymbol{U}}_{gf}, \hat{\boldsymbol{U}}_{ef}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{\sigma}}^2, \hat{\sigma}^2_{gf}, \hat{\sigma}^2_{ef}, \hat{\sigma}^2_{\epsilon}$

Initialization: define $\boldsymbol{K}_m \propto \boldsymbol{G}_m \boldsymbol{W}_m \boldsymbol{G}_m^T$ for each region, and set $\boldsymbol{Z}_{gf} = \boldsymbol{Q}_{gf} \boldsymbol{\Lambda}_{gf}^{\frac{1}{2}}$, $\boldsymbol{Z}_{ef} = \boldsymbol{Q}_{ef} \boldsymbol{\Lambda}_{ef}^{\frac{1}{2}}$, $\boldsymbol{Z}_m = \boldsymbol{Q}_m \boldsymbol{\Lambda}_m^{\frac{1}{2}}$.

while *the increase of ELBO is not negligible* **do**

1: For individual effects: a) update $\boldsymbol{M}_{\boldsymbol{\beta}}$ and $\boldsymbol{S}_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ (equation S1), and b) update $logit(\boldsymbol{\psi})$ for $\boldsymbol{\gamma}$ (equation S2);

2: For cumulative effects: a) update $\boldsymbol{M}_m$ and $\boldsymbol{S}_m$ for $\boldsymbol{U}_m$ (equation S3); b) update $logit(\boldsymbol{\phi})$ for $\boldsymbol{r}$ (equation S4); and c) update a_m and b_m for σ_m^2 (equation S5);

3: For family effects (genetics): a) update $\boldsymbol{M}_{gf}$ and $\boldsymbol{S}_{gf}$ for $\boldsymbol{U}_{gf}$ (equation S6); and b) update a_{gf} and b_{gf} for σ_{gf}^2 (equation S7);

4: For family effects (shared environment): a) update $\boldsymbol{M}_{ef}$ and $\boldsymbol{S}_{ef}$ for $\boldsymbol{U}_{ef}$ (equation S8); and b) update a_{ef} and b_{ef} for σ_{ef}^2 (equation S9);

5: Update a_ϵ and b_ϵ for σ_ϵ^2 according to equation S10;

FIGURE 1
Algorithm 1: Inference procedure using variational Bayes.

Neuroimaging Initiative (ADNI) study ($n = 808$). Pedigree simulator was used to simulate various types of pedigree structures and the gene-dropping method (Huang, Thomas, and Vieland, 2013) was implemented to generate the genotypes of offsprings. Each simulation scenario was replicated 100 times. We randomly split the simulated data into a testing set with 20% samples and a training set with the remaining 80% samples. Pearson correlations and root mean square errors (RMSE) that are calculated based on testing samples were reported for each method.

Scenario 1: The impact of disease model

In this set of simulations, we evaluated the performance of our proposed method under three types of disease models, including outcomes are affected by 1) shared environmental factors only, 2) genetic factors only, and 3) both environmental and genetic factors.

The outcome is affected by shared environmental factors only

To evaluate the impact of shared environmental factors, we randomly selected 3 genes from ADNI dataset and none of them was set to be causal. For simplicity and without loss of generality, we considered mixed two-generation pedigree structures, including a) half-sibling (Supplementary Figure S1A), parents with two offspring (Supplementary Figure S1B) and parents with four offspring (Supplementary Figure S1C). We used 808 samples from ADNI study as founders and formed a total of 394 families including 2040 individuals, which contained 150 individuals from 30 pedigrees of half-siblings, 708 individuals from 177 pedigrees of parents with two offsprings, and 1182 individuals from 197 pedigrees of parents with four offsprings. We simulated the outcomes as $Y_{ij} = \alpha_i + \epsilon_{ij}$, where α_i is the shared environmental effects for family i and $\alpha_i \sim N(0, \sigma_a^2)$. It is straightforward to show that $\mathbf{Y} \sim N(0, \mathbf{K}\sigma_a^2 + \mathbf{I}\sigma^2)$, where $\mathbf{K}$ is a block diagonal matrix with each block being a matrix with all elements equal to 1. Therefore, we simulated the outcomes using $\mathbf{Y} \sim N(0, \mathbf{K}\sigma_a^2 + \mathbf{I}\sigma^2)$, where the percentage of the outcome variance explained by shared environmental factors increased from 25% to 75%.

Pearson correlations and RMSEs are shown in Figure 2. As expected, FBLMM significantly outperformed DPRVB, MultiBLUP and BLMM when shared environmental factors significantly contributed to disease risk. In addition, the prediction accuracy for FBLMM increases as the effects from shared environmental factors increase, but it remains almost unchanged for the other three methods. This is mainly because FBLMM is specifically designed to utilize information from family design for improved prediction. Although adjusting for the relatedness among family members makes it statistically valid to apply population-based methods on family-based studies, overlooking information embedded in the family design can lead to sub-optimal prediction performance. While DPRVB, MultiBLUP and BLMM have similar performance, BLMM tends to be slightly better. This is mainly because BLMM is flexible to the underlying disease models. While MultiBLUP assumes an infinitesimal effect model and DPRVB assumes an isolated effect model, BLMM-based method (i.e., BLMM and FBLMM) can easily accommodate these two commonly used model assumptions.

The outcome is affected by genetic factors only

We evaluated the performance of FBLMM when only genetic variants, including both measured and unmeasured, contributed to the familial aggregation of traits. We first randomly selected three genes and set all of them as causal regions. We simulated the outcomes as $\mathbf{Y} = \sum_{m=1}^{3} \mathbf{g}_m + \boldsymbol{\epsilon}$, where $\mathbf{g}_m$ is the genetic effect for region m and $\mathbf{g}_m \sim N(0, \mathbf{K}_m \sigma_m^2)$. $\mathbf{G}_m$ is an $n \times p_m$ matrix of genetic

10.3389/fgene.2023.1267704

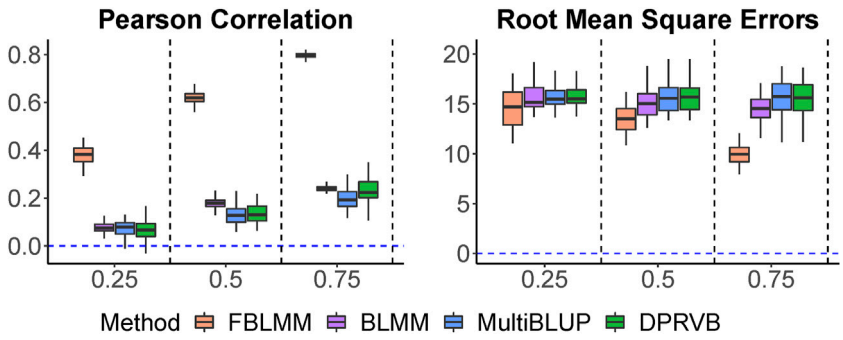

FIGURE 2
The comparison of prediction accuracy when outcomes are only impacted by shared environmental factors. The heritability increases from 25% to 75%.

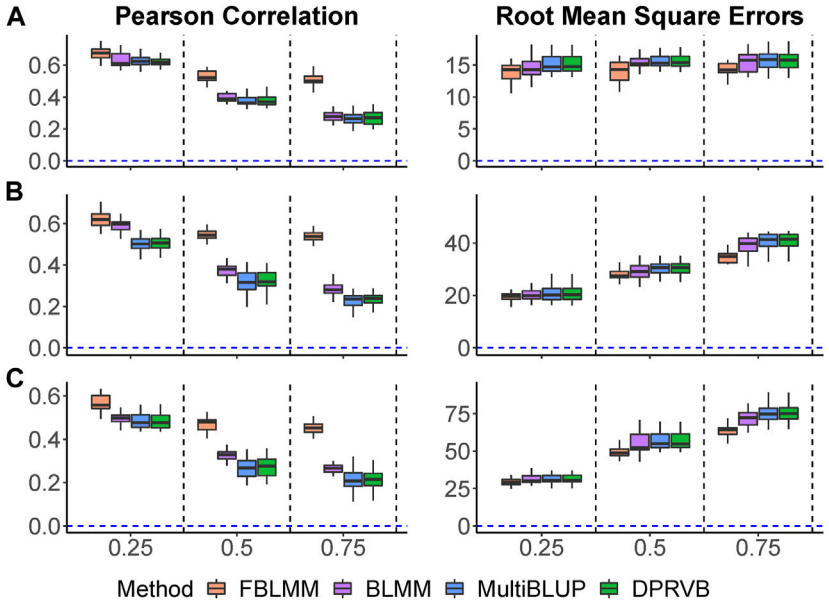

FIGURE 3
The comparison of prediction accuracy when outcomes are affected by unmeasured genetic variants. The total heritability is 60%, and the percentage of heritability accounted by unmeasured variants increases from 25% to 75%. **(A)**: Common variants affect the outcomes ($w_j = 1$). **(B)**: Rare variants affect the outcomes ($w_j = dbeta(MAF_j, 1, 25)^2$). **(C)**: Rare variants affect the outcomes $\left(w_j = \frac{1}{MAF_j(1-MAF_j)}\right)$.

markers on gene m and $\mathbf{K}_m = (\mathbf{G}_m\mathbf{W}_m\mathbf{G}_m^T)/p_m$. Causal genetic variants can be unmeasured in practice (Wen, and Lu, 2017). Therefore, we randomly selected one of the three causal genes as unmeasured (i.e., only two causal genes are in the final simulated dataset). We set the total heritability to be 60% with the proportion of heritability accounted by unmeasured variants changing from 25% to 75%. To evaluate the performance of FBLMM across a range of phenotypes, we first considered the case where outcomes were mainly caused by common variants, and set $w_j = 1$ for each predictor. We then simulated the cases where rare variants contributed substantially to disease risk. We simulated two models under such settings, where a beta-type of weights (denoted as BETA) $w_j = dbeta(MAF_j, 1, 25)^2$ and a weighted sum statistics type of weights were used.

Pearson correlations and RMSEs are shown in Figure 3. As the proportion of genetic variance explained by unmeasured effects increases, the prediction accuracy for all methods decreases with FLBMM decreased the least. For FBLMM, it has robust performance across all settings. When outcomes are mainly caused by common genetic variants (Figure 3. A), FBLMM outperforms the other methods across all simulation settings and captures most of the heritability. This is mainly because FBLMM has an advantage in capturing the genetic effects from unmeasured variants via using the theoretical kinship coefficients. Not surprisingly, the performance of BLMM, MultiBLUP and DPRVB are very similar. When the disease outcomes were simulated under the assumption that rare variants had large contributions (Figure 3. B; Figure 3. C),

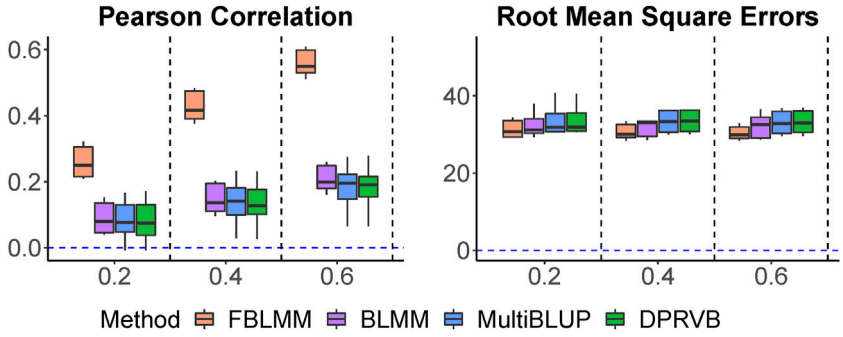

FIGURE 4
The comparison of prediction accuracy when outcomes are affected by both shared environmental factors and genetic variants, including both measured and unmeasured. The totally heritability increases from 20% to 60%, with both genetic and environmental factors contributing equally.

FBLMM performs much better than the existing methods, and BLMM outperforms MultiBLUP and DPRVB. This is mainly because the weights in both FBLMM and BLMM are designed to capture the effects from rare variants. Therefore, FBLMM is expected to have a more robust performance through modeling familial correlations and up-weighting rare genetic variants.

Outcome is affected by shared environmental and genetic factors

In this set of simulations, we evaluated the performance of FBLMM when outcomes were affected by both shared environmental and genetic factors. Three genes were randomly selected as causal, and outcomes were simulated under the following additive model:

$$\mathbf{Y} = \boldsymbol{\alpha} + \sum_{m=1}^{3} \mathbf{g}_m + \boldsymbol{\epsilon}, \tag{9}$$

where $\boldsymbol{\alpha} \sim N(0, \mathbf{K}\sigma_a^2)$ and $\mathbf{K}$ is a block diagonal matrix. $\mathbf{g}_m \sim N(0, \mathbf{K}_m\sigma_m^2)$ and $\mathbf{K}_m = (\mathbf{G}_m\mathbf{G}_m^T)/p_m$. Similar to previous section, among the three causal genes, we randomly set one of them as unmeasured in the data. We gradually increased the percentages of variabilities explained by the shared environmental and genetic effects from 20% to 60%, and both factors contributed equally (i.e., $\sigma_a^2 = \sum_{m=1}^{2}\sigma_m^3$).

Pearson correlations and RMSEs are shown in Figure 4. As the proportion of variability explained by shared environmental and genetic factors increases, the proposed method tends to perform much better than the others. This is because FBLMM is designed to capture predictive effects from both genetic and environmental risk factors simultaneously, whereas the other methods have little ability to model them if they are not measured. Although it is well accepted that family history itself is an important predictor for many complex diseases, little efforts have been made to utilize information embedded in the family design. Our simulation shows that by using the design information, FBLMM can achieve robust performance and substantially improve the prediction models across a range of settings.

Scenario 2: The impact of pedigree structures

In this set of simulations, we assessed the effects of pedigree structures on the performance of FBLMM. We considered the twin design (Supplementary Figure S1D), the trio design (Supplementary Figure S1E), and three-generation pedigree with mixed structures that include 24 avuncular, 30 double cousins, 42 grandparents and 278 sibling (Supplementary Figures S1F–I). We used Eq. 9 to simulate outcomes, where genetic variants on one causal gene is set as unmeasured. Let $\mathbf{g}_u \sim N(0, \mathbf{K}_u\sigma_u^2)$ denote the cumulative predictive effects for the unmeasured gene, and Eq. 9 can be written as $\mathbf{Y} = \boldsymbol{\alpha} + \sum_{m=1}^{2}\mathbf{g}_m + \mathbf{g}_u + \boldsymbol{\epsilon}$.

We considered three types of disease models (Supplementary Table S1: both measured and unmeasured genetic variants have equally contributed to disease risk (i.e., $\sigma_u^2 = \sum_{m=1}^{2}\sigma_m^2$); S_2: shared environmental factors have major influences on disease risk, and measured genetic factors only make small contributions (i.e., $\sigma_a^2 > \sum_{m=1}^{2}\sigma_m^2$); and S_3: both genetic and shared environmental factors were considered with unmeasured genetic variants making major contributions (i.e., $\sigma_u^2 > \sigma_a^2 + \sum_{m=1}^{2}\sigma_m^2$). We set the total heritability for all disease models ranging from 20% to 60%, and the details of parameter settings for each disease model are summarized in Supplementary Table S1.

The results when heritability is 40% are summarized in Figure 5, and the others (i.e., $h = 20\%$ and $h = 60\%$) are shown in Supplementary Figures S2, S3. Under disease model S_1, where measured and unmeasured genetic variants contributed equally to disease risk, Figure 5A showed that the two-generation pedigree design has a higher prediction accuracy as compared to three-generation designs. This is mainly because relatives in two-generation pedigree have higher level of genetic relatedness than those that are far apart. Compared to existing methods, FBLMM worked the best across all pedigree structures under disease model S_1 and captured most of the heritability. Under disease model S_2, where shared environmental factors mainly contributed to disease risk, Figure 5B showed that the existing methods (i.e., BLMM, MultiBLUP and DPR) have lower prediction performance as compared to FBLMM. FBLMM tended to perform similarly across all three pedigree structures considered, as shared environmental factors affect all individuals within the family in a

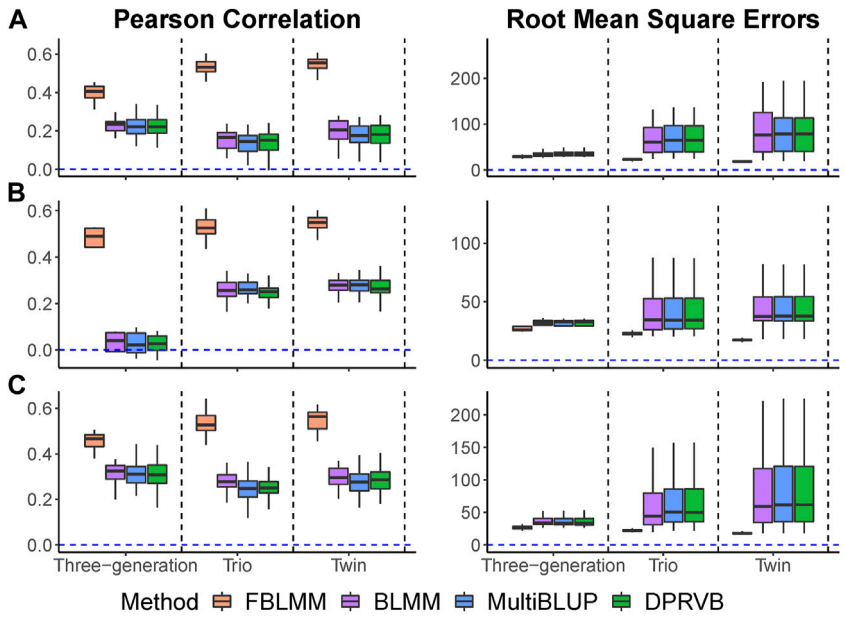

FIGURE 5
The comparison of prediction accuracy under different pedigree structures (h = 40%). Three disease models are considered: **(A)** both measured and unmeasured genetic variants contributed to disease risk; **(B)** shared environmental and measured genetic factors affected outcomes; **(C)** all genetic variants (measured and unmeasured) and shared environmental factors contributed to disease risk.

similar fashion. Under disease model S_3, where both unmeasured and environmental factors contribute significantly to the trait, two-generation pedigree structure tended to have higher prediction accuracy than the three-generation pedigree design (Figure 5C). Regardless of the pedigree structures and disease models considered, our proposed FBLMM always outperformed the other methods (i.e., BLMM, MultiBLUP and DPR). This indicates that FBLMM has robust performance in capturing the predictive effects from shared environmental and unmeasured genetic factors regardless of the pedigree structures. When the heritability is set to be 20% and 60%, the trend remains the same (Supplementary Figures S2, S3). By using the family design information, FBLMM has substantially enhanced the prediction accuracy, and the improvement is robust against various pedigree structures.

Real data application

The proposed method is applied to predict aggressive behavior utilizing the dataset obtained from the Behavioral and Emotional Development in Children (TBED-C) study. TBED-C is a family-based twin study, aimed at discovering genetic factors that contribute to conduct problems in children (Burt and Klump, 2012). TBED-C recruited 1000 twins aged between 6 and 10 years from 500 twin families in Michigan, including 50% monozygous twins. DNA samples were collected from each pair of twins. The sequencing was performed using the Illumina Human Core Exome chip, which includes common variants, rare variants, mitochondrial DNA, and indels. Samples with missing rate $>3\%$ were excluded. Single nucleotide variants (SNVs) were removed if any of the following exclusion criteria was met: 1) call rate $<98\%$ and 2) a p-value for Hardy–Weinberg equilibrium test $<10^{-5}$. After the quality control filtering, there are 957 samples and 513,886 SNVs remained for the analysis. Parents completed the child behavior checklist for each twin separately by rating a series of questionnaires, where children's competencies, behavioral and emotional problems were assessed (Burt and Klump, 2012). Teacher(s) of each twin also completed the report form. Using the recommended approach (Burt and Klump, 2012), we assessed children's aggressive behavior by averaging the raw scale scores from both the parents' and teachers' reports. The distribution of the aggressive scales (AGG, $Mean$ = 3.70; sd = 3.59) is shown in Supplementary Figure S4.

First, to avoid over-fitting and the chance finding problems, 20% samples were randomly select for testing and the rest 80% was used for training. In the training dataset, we assessed the marginal significance for each marker using a linear hybrid model in the GCTA software package (Yang et al., 2011). Common variants with p values >0.1 were filtered out from risk prediction analysis. As a result, approximately 25,168 SNVs remained. This pre-selection aimed to prune a large number of predictors down drastically to a more manageable size, and improve computational speed. We applied all evaluated methods (i.e., FBLMM, BLMM, MulitBLUP and DPRVB) to the remaining genetic variants. Finally, we validated the trained FBLMM model using the test set. The prediction performance was evaluated using Pearson correlation and RMSE. This process was repeated 100 times.

Pearson correlations and the RMSEs are shown in Figure 6. Similar to results from simulations, Figure 6 shows FBLMM performed much better than the others. This clearly indicates that simply adjusting for relatedness among family members can overlook key information, leading to a less accurate risk prediction model. On contrary, utilizing information embedded in the family

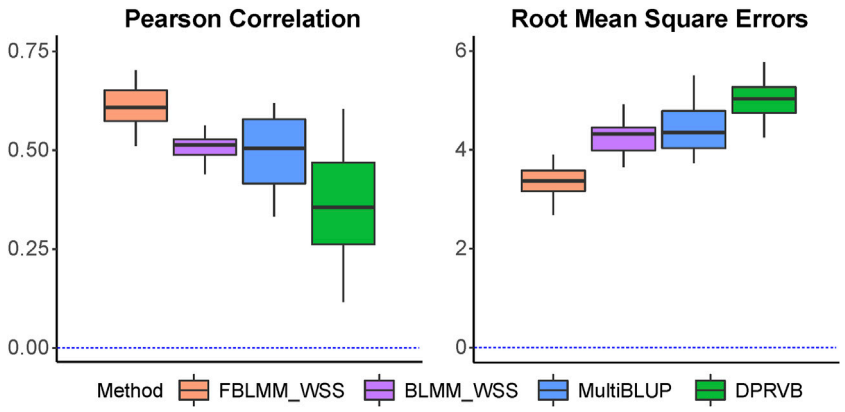

FIGURE 6
The comparison of prediction accuracy for aggressive scale.

design can substantially improve prediction accuracy, as this makes the model more flexible to capture the predictive effects from unknown genetic and shared environmental risk factors.

Discussion and conclusions

In this paper, we have developed a novel FBLMM method for risk prediction analysis on sequencing data obtained from family-based genetic studies. Fundamentally different from existing methods that adjust for family correlations, FBLMM utilizes this relatedness to further improve prediction accuracy. Specifically, it forms two surrogates, including a theoretical kinship coefficient matrix (i.e., $\mathbf{K}^{g_f}$) and a block diagonal matrix (i.e., $\mathbf{K}^{e_f}$), to capture effects from unmeasured genetic and shared environmental factors. In addition, FBLMM extends the BLMM method proposed by (Hai and Wen, 2020), and thus it inherits all the advantages in the BLMM method. For example, it infers its parameters using variational Bayes algorithm rather than the traditional MCMC, making it much more computationally efficient. Supplementary Table S2 provided the details of computational resources needed as the sample size and the number of variants increase. Furthermore, it can capture predictive effects from both common and rare variants, and easily accommodate various model assumptions (e.g., isolated large effects and infinitesimal model). It is worth noting that although we mainly focused on genetic variants, our proposed framework has the intrinsic capacity in modeling the predictive effects from important demographic variables, where their predictive effects can be selected and modelled through the fixed effects (i.e., $\mathbf{X\beta}$) in our model. For example, in addition to genetic information, we can add family history, age and gender into the fixed effect part (i.e., $\mathbf{X}$) of our model, and their predictive effects can be directly estimated by our proposed framework. Through simulation studies, we have shown that FBLMM can yield higher prediction accuracy than existing methods, and our analysis on Michigan Twin data has also showed that FBLMM can better predict AGG.

The importance of genetic and environmental factors in risk prediction has long been appreciated (Nilsson et al., 2004). Many previous studies have shown that a substantial amount of heritability can be explained by family information due to a combination of genetic factors and shared environmental conditions (Bermejo and Hemminki, 2005; Gim et al., 2017). The family information can be helpful in identifying sub-populations that are at high risk (MacInnis et al., 2011; So et al., 2011; Gim et al., 2017). Despite its clinical importance, few methods fully use this information when building risk prediction models based on high-dimensional genomic data obtained from family-based studies. Existing analytical methods are usually an extension of the models designed for population-based studies, and thus they tend to make the observations un-correlated before estimating the predictive effects from genetic variants (Meigs et al., 2008). While this most common practice can allow researchers to build a statistically valid risk prediction model using genomic data from family-based study designs, it overlooks important information embedded in the design, leading to a model with decreased prediction accuracy. In this study, one of the key features of our proposed model is that it utilizes the family design to improve prediction model, rather than simply adjusting for the correlations among family members. Based on the design information, we formed two surrogate measures, including a theoretical kinship coefficient matrix (i.e., $\mathbf{K}^{g_f}$) and a block diagonal matrix (i.e., $\mathbf{K}^{e_f}$), to capture the impacts of genetic and environmental risk factors. As shown in our simulation studies (Figure 2 to Figure 5) and the analysis of TBED-C dataset (Figure 6), we have shown that FBLMM have outperformed commonly used methods via using the design information, indicating our proposed method has the capacity to substantially improve prediction models for family-based studies.

Rare variants of large effects can play an important role in complex human diseases (Gaukrodger et al., 2005). It has been reported that the largest contributions to genetic risk of human diseases can come from rare variants (Mancuso et al., 2016; Hernandez et al., 2019). However, few family-based genetic studies are powerful enough to model these effects, primarily due to the lack of efficient analytical methods (McIntosh et al., 2016). We have recently developed BLMM for risk prediction

studies using genomic data from population-based study designs (Hai and Wen, 2020), and BLMM has achieved an improved prediction accuracy through simultaneously considering both common and rare variants. Instead of modeling individual predictive effects that are hard to estimate for rare variants, BLMM models the cumulative predictive effects from a group of variables that include both common and rare variants. BLMM uses a WSS weight function that has been used in association analysis of sequencing data to address the contributions of rare variants (Wu et al., 2011), and this leads to an improvement for prediction studies. Our proposed FBLMM is built within the BLMM framework, and thus it inherits BLMM's capacity in modeling rare variants. Same as BLMM, FBLMM uses the WSS function to up-weight the rare variants so that their predictive effects can be effectively captured. As shown in simulations, FBLMM can achieve better assessment, when outcomes were simulated under the assumption that rare variants significantly contribute to the risk (Figures 2B, C).

One of the limitations of our method is that it overlooks the contributions of non-additive effects, especially interactions. As indicated in existing literature (Weissbrod, Geiger, and Rosset, 2016), non-linear predictive effects (e.g., epistasis) widely exist. Therefore, it is important to incorporate non-additive effects into risk prediction models. A potential solution within the FBLMM framework is to kernelize the variance-covariance matrix of the random effect terms, so that the assumed relationships between predictors and outcomes can be non-linear. For example, similar to MKLMM (Weissbrod, Geiger, and Rosset, 2016), polynomial kernel of two degrees of freedom and the saturating pathway kernel can be used to capture non-linear predictive effects. This will be a future direction of our research.

In summary, we have proposed a Bayesian linear mixed model for risk prediction analysis on genomic data obtained from family-based study designs. Our proposed FBLMM extends the BLMM method, and thus it can not only capture the predictive effects from both common and rare variants, but also accommodate various disease model assumptions. In addition, using study design information, FBLMM forms two surrogates to model the predictive effects from unmeasured/unknown genetic and environmental risk factors, which substantially facilitates family-based prediction studies. The algorithm implementing our proposed method is available at https://github.com/yhai943/FBLMM.

Data availability statement

The R package is available at https://github.com/yhai943/FBLMM. The data presented in the study are deposited at figshare repository (https://figshare.com/s/02c32e50c5b7529c0fb5). The use of the original genotype data is subject to the approval from the TBED-C study team (https://msutwinstudies.com/msutr-data).

Author contributions

YH: Conceptualization, Methodology, Formal Analysis, Visualization, Writing–original draft. WZ: Writing–original draft, Software. QM: Formal Analysis. LL: Supervision. YW: Conceptualization, Methodology, Writing–review and editing, Project administration, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project is funded by the National Natural Science Foundation of China (Award No. 82173632), the Early Career Research Excellence Award from the University of Auckland, and the Marsden Fund from Royal Society of New Zealand (Project No. 19-UOA-209).

Acknowledgments

The author(s) 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. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation and Employment's Research Infrastructure programme.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1267704/full#supplementary-material

References

Bermejo, J. L., and Hemminki, K. (2005). Familial lung cancer and aggregation of smoking habits: A simulation of the effect of shared environmental factors on the familial risk of cancer. *Cancer Epidemiol. Prev. Biomarkers* 14 (7), 1738–1740. doi:10.1158/1055-9965.EPI-05-0201

Burt, S. A., and Klump, K. L. (2012). Etiological distinctions between aggressive and non-aggressive antisocial behavior: results from a nuclear twin family model. *J. Abnorm. Child Psychol.* 40 (7), 1059–1071. doi:10.1007/s10802-012-9632-9

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.* 103 (484), 1438–1456. doi:10.1198/016214508000000869

Chen, C.Y.-C., Scurrah, K. J., Stankovich, J., Garoufalis, P., Dirani, M., Per-tile, K. K., et al. (2007). Heritability and shared environment estimates for myopia and associated ocular biometric traits: the genes in myopia (gem) family study. *Hum. Genet.* 121 (3-4), 511–520. doi:10.1007/s00439-006-0312-0

Chen, Z., and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* 59 (4), 762–769. doi:10.1111/j.0006-341x.2003.00089.x

Couillard, C., Despr´es, J.-P., Lamarche, B., Bergeron, J., Gagnon, J., Leon, A. S., et al. (2001). Effects of endurance exer-cise training on plasma HDL cholesterol levels depend on levels of triglycerides: evidence from men of the health, risk factors, exercise training and genetics (heritage) family study. *Arteriosclerosis, thrombosis, Vasc. Biol.* 21 (7), 1226–1232. doi:10.1161/hq0701.092137

Cruceanu, C., Ambalavanan, A., Spiegelman, D., Gauthier, J., Lafreni`ere, R. G., Dion, P. A., et al. (2013). Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. *Genome* 56 (10), 634–640. doi:10.1139/gen-2013-0081

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare vari-ants create synthetic genome-wide associations. *PLoS Biol.* 8 (1), 1000294. doi:10.1371/journal.pbio.1000294

Dirani, M., Chamberlain, M., Shekar, S. N., Islam, A. F., Garoufalis, P., Chen, C. Y., et al. (2006). Heritability of refractive error and ocular bio-metrics: the genes in myopia (gem) twin study. *Investigative Ophthalmol. Vis. Sci.* 47 (11), 4756–4761. doi:10.1167/iovs.06-0270

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, Brazil, September 2017, 328–332.

Gaukrodger, N., Mayosi, B., Imrie, H., Avery, P., Baker, M., Connell, J., et al. (2005). A rare variant of the leptin gene has large effects on blood pressure and carotid intima-medial thickness: A study of 1428 individuals in 248 families. *J. Med. Genet.* 42 (6), 474–478. doi:10.1136/jmg.2004.027631

Gim, J., Kim, W., Kwak, S. H., Choi, H., Park, C., Park, K. S., et al. (2017). Improving disease prediction by incorporating family disease history in risk prediction models with large-scale genetic data. *Genetics* 207 (3), 1147–1155. doi:10.1534/genetics.117.300283

Hai, Y., and Wen, Y. (2020). A Bayesian linear mixed model for prediction of complex traits. *Bioinformatics* 36 (22-23), 5415–5423. doi:10.1093/bioinformatics/btaa1023

Helgadottir, A., Gretarsdottir, S., Thorleifsson, G., Hjartarson, E., Sigurdsson, A., Magnusdottir, A., et al. (2016). Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat. Genet.* 48 (6), 634–639. doi:10.1038/ng.3561

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* 51 (9), 1349–1355. doi:10.1038/s41588-019-0487-7

Huang, Y., Thomas, A., and Vieland, V. J. (2013). Employing MCMC under the PPL frame-work to analyze sequence data in large pedigrees. *Front. Genet.* 4, 59. doi:10.3389/fgene.2013.00059

Ionita-Laza, I., and Ottman, R. (2011). Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 189 (3), 1061–1068. doi:10.1534/genetics.111.131813

Ji, W., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., et al. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.* 40 (5), 592–599. doi:10.1038/ng.118

Laitinen, T., Rasanen, M., Kaprio, J., Koskenvuo, M., and Laitinen, L. A. (1998). Importance of genetic factors in adolescent asthma: A population-based twin-family study. *Am. J. Respir. Crit. care Med.* 157 (4), 1073–1078. doi:10.1164/ajrccm.157.4.9704041

Lali, R., Chong, M., Omidi, A., Mohammadi-Shemirani, P., Le, A., and Pare, G. (2020). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *bioRxiv*. doi:10.1164/ajrccm.157.4.9704041

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91 (2), 224–237. doi:10.1016/j.ajhg.2012.06.007

Lichtenstein, P., Yip, B. H., Bj¨ork, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., et al. (2009). Common genetic determinants of schizophrenia and bipolar disor-der in Swedish families: A population-based study. *Lancet* 373 (9659), 234–239. doi:10.1016/S0140-6736(09)60072-6

MacInnis, R. J., Antoniou, A. C., Eeles, R. A., Severi, G., Al Olama, A. A., McGuf-fog, L., et al. (2011). A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. *Genet. Epidemiol.* 35 (6), 549–556. doi:10.1002/gepi.20605

Mancuso, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., Quinque, D., et al. (2016). The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* 48 (1), 30–35. doi:10.1038/ng.3446

Marateb, H. R., Mohebian, M. R., Javanmard, S. H., Tavallaei, A. A., Tajadini, M. H., Heidari-Beni, M., et al. (2018). Prediction of dyslipidemia using gene mutations, family history of diseases and anthropometric indicators in children and adolescents: the caspian-iii study. *Comput. Struct. Biotechnol. J.* 16, 121–130. doi:10.1016/j.csbj.2018.02.009

McIntosh, A. M., Hall, L. S., Zeng, Y., Adams, M. J., Gibson, J., Wigmore, E., et al. (2016). Genetic and environmental risk for chronic pain and the contribution of risk variants for major depressive disorder: A family-based mixed-model analysis. *PLoS Med.* 13 (8), 1002090. doi:10.1371/journal.pmed.1002090

Meigs, J. B., Shrader, P., Sullivan, L. M., McAteer, J. B., Fox, C. S., Dupuis, J., et al. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* 359 (21), 2208–2219. doi:10.1056/NEJMoa0804742

Mihaescu, R., Pencina, M. J., Alonso, A., Lunetta, K. L., Heckbert, S. R., Ben-jamin, E. J., et al. (2013). Incremental value of rare genetic variants for the prediction of multifactorial diseases. *Genome Med.* 5 (8), 76. doi:10.1186/gm480

Nilsson, E., Salonen Ros, H., Cnattingius, S., and Lichtenstein, P. (2004). The importance of genetic and environmental effects for pre-eclampsia and gestational hypertension: A family study. *BJOG Int. J. Obstetrics Gynaecol.* 111 (3), 200–206. doi:10.1111/j.1471-0528.2004.00042x.x

Peloso, G. M., Auer, P. L., Bis, J. C., Voorman, A., Morrison, A. C., Stitziel, N. O., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* 94 (2), 223–232. doi:10.1016/j.ajhg.2014.01.009

Ruderfer, D. M., Korn, J., and Purcell, S. M. (2010). Family-based genetic risk prediction of multifactorial disease. *Genome Med.* 2 (1), 2. doi:10.1186/gm123

Ramachandrappa, S., Raimondo, A., Cali, A. M., Keogh, J. M., Henning, E., Saeed, S., et al. (2013). Rare variants in single-minded 1 (SIM1) are associated with severe obesity. *J. Clin. investigation* 123 (7), 3042–3050. doi:10.1172/JCI68016

So, H.-C., Kwan, J. S., Cherny, S. S., and Sham, P. C. (2011). Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* 88 (5), 548–565. doi:10.1016/j.ajhg.2011.04.001

Speed, D., and Balding, D. J. (2014). MultiBLUP: improved snp-based prediction for complex traits. *Genome Res.* 24 (9), 1550–1557. doi:10.1101/gr.169375.113

Stefansson, H., Rujescu, D., Cichon, S., Pietil¨ainen, O. P., Ingason, A., Steinberg, S., et al. (2008). Large recurrent microdeletions associated with schizophrenia. *nature* 455 (7210), 232–236. doi:10.1038/nature07229

Valdez, R., Greenlund, K. J., Khoury, M. J., and Yoon, P. W. (2007). Is family history a useful tool for detecting children at risk for diabetes and cardiovascular diseases? A public health perspective. *Pediatrics* 120, 78–86. SUPPLEMENT 2. doi:10.1542/peds.2007-1010G

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi:10.3168/jds.2007-0980

Wang, S., Yang, Z., Ma, J. Z., Payne, T. J., and Li, M. D. (2014). Introduction to deep sequenc-ing and its application to drug addiction research with a focus on rare variants. *Mol. Neurobiol.* 49 (1), 601–614. doi:10.1007/s12035-013-8541-4

Weissbrod, O., Geiger, D., and Rosset, S. (2016). Multikernel linear mixed models for complex phenotype prediction. *Genome Res.* 26 (7), 969–979. doi:10.1101/gr.201996.115

Wen, Y., and Lu, Q. (2017). Risk prediction modeling on family-based sequencing data using a random field method. *Genetics* 117. doi:10.1534/genetics.117.199752

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant associa-tion testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89 (1), 82–93. doi:10.1016/j.ajhg.2011.05.029

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi:10.1016/j.ajhg.2010.11.011

Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* 8 (1), 456. doi:10.1038/s41467-017-00470-2

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* 9 (2), 1003264. doi:10.1371/journal.pgen.1003264