Bioinformatics, 36(6), 2020, 1785–1794
doi: 10.1093/bioinformatics/btz822
Advance Access Publication Date: 6 November 2019
Original Paper

OXFORD

Genetics and population analysis

Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data

Jun Li[1], Qing Lu[2] and Yalu Wen [3,*]

[1]Department of Thoracic Surgery, Dalian Municipal Central Hospital Affiliated of Dalian Medical University, Dalian 116000, China, [2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI 48824, USA and [3]Department of Statistics, University of Auckland, Auckland 1010, New Zealand

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on June 10, 2019; revised on October 8, 2019; editorial decision on October 29, 2019; accepted on November 1, 2019

Abstract

Motivation: The use of human genome discoveries and other established factors to build an accurate risk prediction model is an essential step toward precision medicine. While multi-layer high-dimensional omics data provide unprecedented data resources for prediction studies, their corresponding analytical methods are much less developed.
Results: We present a multi-kernel penalized linear mixed model with adaptive lasso (MKpLMM), a predictive modeling framework that extends the standard linear mixed models widely used in genomic risk prediction, for multi-omics data analysis. MKpLMM can capture not only the predictive effects from each layer of omics data but also their interactions via using multiple kernel functions. It adopts a data-driven approach to select predictive regions as well as predictive layers of omics data, and achieves robust selection performance. Through extensive simulation studies, the analyses of PET-imaging outcomes from the Alzheimer's Disease Neuroimaging Initiative study, and the analyses of 64 drug responses, we demonstrate that MKpLMM consistently outperforms competing methods in phenotype prediction.
Availability and implementation: The R-package is available at https://github.com/YaluWen/OmicPred.
Contact: y.wen@auckland.ac.nz
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The use of human genome discoveries and other established risk factors for predicting disease risk is an essential step toward precision medicine, an emerging model of healthcare that tailors treatment strategies based on individuals' profiles (Ashley, 2015). The advert of high-throughput multi-platform genomic technologies has led to the accumulation of diverse types of molecular data (e.g. genome, transcriptome, methylome and proteome data) (Boekel *et al.*, 2015). Each type provides a unique, complementary and partially independent view of the disease mechanisms, and thus embeds the essential information for building an accurate risk prediction model. Therefore, an integrative framework that simultaneously utilizes multi-omics data can substantially deepen our understanding of hereditary and environmental causes in disease etiology (Ritchie *et al.*, 2015).

In the existing literature, many strategies have emerged to integrate multi-omics data for association studies (Bersanelli *et al.*, 2016; Morris and Baladandayuthapani, 2017; Zeng and Lumley, 2018). The main objectives of these studies usually are to understand the inter-relationships among various omics data and to detect phenotype-related modules (Bersanelli *et al.*, 2016). For example, unsupervised matrix factorization methods [e.g. iCluster (Shen *et al.*, 2009) and JIVE (Lock *et al.*, 2013)] project the variations among multi-omics data on to a dimension-reduced space that includes a common basis space capturing the coherent patterns across all data types. Correlation-based methods [e.g. canonical correlation analysis (Lin *et al.*, 2013b) and partial least squares (Chen and Zhang, 2016)] maximize the correlations between multi-omics data and find the fundamental relationships among them. Mixed graphical models and Bayesian methods [e.g. FuseNet (Zitnik and Zupan, 2015) and Prob_GBM (Cho and Przytycka, 2013)], which make explicit assumptions on the distributions of various data types and the dependency structures, build a probabilistic framework to model and estimate relationships among multi-omics data. Multi-step (or multi-stage) methods, which find relationships between multi-omics data first and then between multi-omics and the phenotype of interest, are commonly used strategy to detect phenotype-related modules. For example, SNF (Wang *et al.*, 2014) first integrates different data types by constructing a network for each data type, and then fuses these

networks into one comprehensive network that is further used for patient classifications. rMKL-LPP (Speicher and Pfeifer, 2015) first uses a multiple kernel learning algorithm, where the multi-omics data are projected to a lower dimensional and integrative subspace, and then clusters patients using *k*-means based on the distance derived from the multi-kernel learning algorithm. While the existing methods offer important advances for multi-omics data analyses, they are typically designed for the discovery of molecular mechanisms and sample classifications. Therefore, they are not directly applicable for predicting the disease outcomes, especially for continuous outcomes.

Most common diseases, such as type 2 diabetes and cancers, are affected by multiple variants through complex biological pathways, and thus progress toward accurately predicting phenotypes requires the advanced analytical methods that can model all variants from various molecular levels jointly (Morris and Baladandayuthapani, 2017). Linear mixed models (LMMs), the method of choice for prediction modeling with single-layer genomic data, have considerable potential to be extended for multi-omics data (Speed and Balding, 2014; VanRaden, 2008; Weissbrod *et al.*, 2016; Yang *et al.*, 2010, 2011). At the core, LMMs assume that genetically similar individuals have similar outcomes, and relate the outcomes with genetic similarity matrix (GSM) through specifying a random effect for each GSM. GSM, which is used to encode various types of genetic effects, is traditionally estimated by theoretical kinship coefficients but nowadays determined by variants from genome-wide data (Speed and Balding, 2014; VanRaden, 2008; Weissbrod *et al.*, 2016; Yang *et al.*, 2010). For example, genomic best linear unbiased prediction uses a random effect term in LMM to model the disease outcome, where the correlation structure is specified according to the GSM estimated directly from whole-genome data (VanRaden, 2008). The recently developed MultiBLUP uses multiple random effects to allow for different effect sizes of genetic variants located at various genomic regions (e.g. coding and eQTLs), where the covariance structure for each random effect is determined by the GSM estimated from the genetic variants within the region (Speed and Balding, 2014). Multi-kernel linear mixed model (MKLMM) generalizes MultiBLUP by using kernel functions under the reproducing kernel Hilbert space to estimate GSMs for each genomic region so that the nonlinear effects within each genomic region can be considered (Weissbrod *et al.*, 2016).

LMM-based methods encode genetic effects from multiple variants through the GSMs, which substantially reduces the dimension of genomic data and makes it appealing for high-dimensional genomic data analyses. Similar idea can be adopted for modeling high-dimensional multi-omics data, where the genetic similarities are replaced by omics-similarities. Though promising, there are several challenges for directly adopting the LMM framework for risk prediction analysis on multi-omics data. Multi-omics data with each omics data being high-dimensional have a huge amount of noise. Utilizing omics-similarities that are constructed based on all genomic regions and all available omics data can substantially attenuate the effects of predictive regions, and thus reduce the robustness and accuracy of the prediction model. The underlying biological rules can introduce a hierarchical structure to the measured variables from multi-omics data (Morris and Baladandayuthapani, 2017). For example, methylation can change the activity of a DNA segment, and it typically represses the gene expression when located at a gene promoter region. A recent study showed that there is strong evidence for interaction between genotype and methylation on change in triglycerides (Fisher *et al.*, 2018). The complex intra/inter-relationships among multi-omics data and their dependent effects on the disease outcome can raise modeling and inferential challenges (Zeng and Lumley, 2018). The existing LMMs that usually assume the effects of genetic variants are independent and thus cannot directly be applied to model multi-omics data. Furthermore, interactions widely exist and it is crucial to capture the potential interaction effects while building prediction models (Buil *et al.*, 2015; Moore and Williams, 2009). While most of the existing LMMs focus on additive effects, few have investigated the high-order interactions from genome-wide data (Weissbrod *et al.*, 2016). Though the recent proposed MKLMM has the potential to model the high-order interactions by using kernel functions under RKHS, it only uses one kernel (e.g. linear or saturate pathway kernels) per region and thus only accounts for one specific effect. As the genetic architecture of complex diseases is unknown in advance and different omics data types may have different kinds of effects on the disease outcome, it is suboptimal to use only one kernel function to capture the predictive effects from all layers of omics data. A data-adaptive algorithm that can choose multiple kernel functions from the data to capture various types of effects is needed.

To address these challenges, we develop a multi-kernel penalized linear mixed model (MKpLMM) for prediction analysis on high-dimensional multi-omics data. The MKpLMM (i) uses multiple random effects to allow for heterogeneous effect size distributions from different regions, (ii) adaptively selects multiple appropriate kernel functions for each predictive region and accounts for potential interactions among multi-omics data and (iii) provides the theoretical justification for the selections of predictive regions (i.e. selecting random effects) and predictive variants (i.e. selecting fixed effects) under high-dimensional settings. The proposed MKpLMM is a flexible framework for risk prediction analyses on multi-omics data, where nonlinear effects among multi-omics data can be modeled and predictive regions can be efficiently selected. In the following sections, we will lay out the details of the MKpLMM method and its theoretical properties. We will compare its accuracy with other methods, and illustrate it through applications to datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Saykin *et al.*, 2010) and the drug responses in chronic lymphocytic leukemia (CLL) patients study (Dietrich *et al.*, 2017).

2 Materials and methods

MKpLMM extends LMMs used for genomic data analyses via kernelization of its covariance matrix and penalizing the log-likelihood function to obtain sparse solutions. In the following sections, we will first briefly overview LMMs and then present our model for multi-omics data analysis. Finally, we will present a kernel principle component analysis to deal with the potential correlations.

2.1 Prediction with linear mixed model

For the prediction modeling with genomic data, LMM assumes that the outcome vector of n individual (denoted by $\boldsymbol{Y}$) is influenced by demographic variables (denoted by $\boldsymbol{X}_d$), genetic effects from the rth region (denoted by $\boldsymbol{g}^r$) and a random error (denoted by ϵ) via

$$\boldsymbol{Y} = \boldsymbol{X}_d\boldsymbol{\beta}_d + \sum_r^R \boldsymbol{g}^r + \epsilon, \tag{1}$$

where $\boldsymbol{g}^r \sim N(0, \boldsymbol{K}^r\sigma_r^2)$ and $\epsilon \sim N(0, \boldsymbol{I}_n\sigma_e^2)$. $\boldsymbol{K}^r$ is the GSM for the rth region, $\boldsymbol{\beta}_d$ is the effects of demographic variables (e.g. age and gender), and R is the total number of regions considered. A commonly used GSM for genetic data is $\boldsymbol{K}^r = \boldsymbol{G}_r\boldsymbol{G}_r^T/p_r$, where $\boldsymbol{G}_r$ is a $n \times p_r$ matrix of genotypes located on region r. If this holds for every regions, then Equation (1) can be written as, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_d + \sum_r^R \sum_j^{p_r} G_{rj}\gamma_{rj} + \epsilon$, where $\gamma_{rj} \sim N(0, \sigma_r^2/p_r)$, G_{rj} is the jth column of $\boldsymbol{G}_r$, and γ_{rj} is its effect size.

Based on a similar idea, for multi-omics data modeling, we assume the outcome vector can be modeled as a sum of omics' effects (denoted by $\boldsymbol{O}_r$) from predictive genomic regions. We further assume that the omics' effects from each region can be decomposed into omic-specific effect and their corresponding interactions. Here we use gene boundary to define regions, but other criteria (e.g. pathway) can also be used to define regions. The general form of our model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}_d\boldsymbol{\beta}_d + \sum_r^R \boldsymbol{O}^r + \epsilon = \boldsymbol{\beta} + \sum_r^R \sum_{j\in O_r} \boldsymbol{o}_j^r + \epsilon, \tag{2}$$

where $\boldsymbol{o}_j^r \sim N(0, \boldsymbol{K}_j^r\sigma_{rj}^2)$ and $\epsilon \sim N(0, \boldsymbol{I}_n\sigma_0^2)$. We use $\boldsymbol{X} = (\boldsymbol{X}_d, \boldsymbol{X}_o)$ to denote fixed effect variables with $\boldsymbol{X}_d$ being a $n \times p_d$ dimensional

demographic variables and $\boldsymbol{X_o}$ being a $n \times p_o$ dimensional predictors from omics data that are treated as fixed effects. $\boldsymbol{\beta} = (\boldsymbol{\beta}_d^T, \boldsymbol{\beta}_o^T)^T$ are fixed effects with $\boldsymbol{\beta_d}$ being a $p_d \times 1$ vector of fixed effects for demographic variables and $\boldsymbol{\beta_o}$ being a $p_o \times 1$ vector of fixed effects for omics predictors. O_r is the set of all omics-effects including potential interaction effects (e.g. genetic effect, methylation effect and interaction effect between genetic variants and methylation levels) considered in region r, and $\boldsymbol{o}_j^r$ is the jth omics effect from region r.

For example, we consider genomic, methylation and gene expression data. We define regions based on gene, and assume that gene expression is an outcome of the joint regularization of genetic and epigenetic effects. We treat the predictive effects from gene expression data as fixed effects, and only consider interactions between methylation and genomic data. Model (2) can be written as,

$$\boldsymbol{Y} = \boldsymbol{X_d\beta_d} + \boldsymbol{X_o\beta_o} + \sum_r^R \sum_{j\in(g,m,gm)} \boldsymbol{o}_j^r + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{X_o}$ is a $n \times p_o$ dimensional matrix of gene expression data and $\boldsymbol{\beta_o}$ is their effects. $\boldsymbol{o}_g^r \sim N(0, \boldsymbol{K}_g^r\sigma_{rg}^2)$, $\boldsymbol{o}_m^r \sim N(0, \boldsymbol{K}_m^r\sigma_{rm}^2)$ and $\boldsymbol{o}_{gm}^r \sim N(0, \boldsymbol{K}_{gm}^r\sigma_{rgm}^2)$ represent the genetic effect, methylation effect and the interaction effect between genetic variants and methylation levels at each CpG site located on region r, respectively. $\boldsymbol{K}_g^r$, $\boldsymbol{K}_m^r$ and $\boldsymbol{K}_{gm}^r$ respectively denote the similarity in genetic variants, methylation levels and their interaction for region r. If linear kernel is used for both genetic data and methylation data, then $\boldsymbol{K}_g^r = \boldsymbol{G}_r\boldsymbol{G}_r^T/p_{rg}$ and $\boldsymbol{K}_m^r = \boldsymbol{M}_r\boldsymbol{M}_r^T/p_{rm}$, with $\boldsymbol{G}_r(\boldsymbol{M}_r)$ being a $n \times p_{rg}$ ($n \times p_{rm}$) matrix denoting the genotypes (methylation levels) for region r and p_{rg} (p_{rm}) being the number of genetic variants (CpG sites) for region r. For simplicity, here we only consider pairwise interaction effects and define $\boldsymbol{K}_{gm}^r$ as $\boldsymbol{K}_{gm}^r = \boldsymbol{K}_g^r \circ \boldsymbol{K}_m^r$ with $\circ$ being the Hadamard product, then Equation (3) is equivalent to

$$\begin{aligned}\boldsymbol{Y} = \boldsymbol{X\beta}_d + \sum_r^R \boldsymbol{X}_{or}\beta_r^e + \sum_r^R\sum_j^{p_{rg}} G_{rj}\beta_{rj}^g + \sum_r^R\sum_j^{p_{rm}} M_{rj}\beta_{rj}^m \\ + \sum_r^R\sum_l^{p_{rm}}\sum_j^{p_{rg}} M_{rl}G_{rj}\beta_{rlj}^f + \boldsymbol{\varepsilon},\end{aligned}$$

where β_r^e is the gene expression effect, $\beta_{rj}^g \sim N(0, \sigma_{rg}^2/p_{rg})$ is the effect of the jth genetic variant from the rth region, $\beta_{rl}^m \sim N(0, \sigma_{rm}^2/p_{rm})$ is the effect of methylation level on lth CpG site from the rth region and $\beta_{rlj}^f \sim N(0, \sigma_{rf}^2/p_{rf})$ is their pairwise interaction effect.

The proposed model (2) can model more complicated settings. For example, the region can be defined based on pathways and we wish to consider pairwise and three-way interactions among all omics data, $\boldsymbol{Y} = \boldsymbol{X_d\beta_d} + \sum_r^R \sum_{j\in O_r} \boldsymbol{o}_j^r + \epsilon$, $\boldsymbol{o}_j^r \sim N(0, \boldsymbol{K}_j^r\sigma_{rj}^2)$, where $O_r = (e, g, m, eg, em, gm, egm)$. $\boldsymbol{o}_e^r$, $\boldsymbol{o}_g^r$ and $\boldsymbol{o}_m^r$ respectively represent the transcriptomic, genomic, methylation effects. $\boldsymbol{o}_{eg}^r$, $\boldsymbol{o}_{em}^r$, $\boldsymbol{o}_{gm}^r$ respectively denote the interaction effects between genetic variants and gene expression levels, methylation and transcriptomic levels, as well as genetic variants and methylation levels. $\boldsymbol{o}_{egm}^r$ is the interaction effect among gene expression levels, genetic variants and methylation levels from the rth region.

Supposed that phenotypes are measured for individuals indexed by S and we want to make predictions for those individuals in the set T, where phenotypes are unknown. Given the parameter estimates, the predicted values can be obtained as $\hat{Y}_T = X_T\hat{\boldsymbol{\beta}} + \sum_r^R \sum_{j\in O_r} \hat{o}_j^{rT}$, where $\hat{o}_j^{rT} = \mathrm{K}_{\mathrm{j}}^{\mathrm{r}}(\mathrm{K}_{\mathrm{jSS}}^{\mathrm{r}})^{-1}\hat{o}_j^{rS}$, $\mathrm{K}_{\mathrm{jTS}}^{\mathrm{r}}$ and $\boldsymbol{K}_{jSS}^r$ are sub-matrices of $\boldsymbol{K}_j^r$ defined by the subscripts, and $\hat{o}_j^{rS}$ are estimated from model (2).

2.2 Penalized maximum likelihood estimator

The underlying causes for many complex diseases are unknown in advance, and thus it is quite likely that a substantial amount of regions included in the analyses are not predictive, especially for high-dimensional data. Moreover, it is possible only some layers of omics data are predictive. Therefore, including all omics layers and their possible interactions in the analyses can attenuate the effects of predictors and result in sub-optimum performance. Variable selection is of great importance for prediction analyses on high-dimensional multi-layer omics data (Byrnes *et al.*, 2013; Morris and Baladandayuthapani, 2017). As the proposed LMM framework is very flexible and can be easily adapted to various situations, its variable selection procedure involves both fixed and random effects. For example, for model (3), the selection of genes whose expression levels are predictive involves the selection of fixed effect (i.e. $\boldsymbol{\beta}_o = 0$), and the selection of CpG sites and genetic variants requires the selection of random effects (i.e. $\sigma_{rg}^2 = 0$, $\sigma_{rm}^2 = 0$, and $\sigma_{rgm}^2 = 0$).

A natural choice for simultaneously selecting and estimating parameters from LMMs is to add a L_1 type penalty to the log-likelihood function. Under model (2), we use $\boldsymbol{\theta}_r = \cup_{j\in O_r}(\sigma_{rj}^2)$ to denote all the parameters for random effects in region r and $\boldsymbol{\theta}_{\text{omics}} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_{\mathrm{R}}^T)^T$ to denote all parameters related to random effects. Let $\boldsymbol{\theta} = (\sigma_0^2, \boldsymbol{\theta}_{\text{omics}}^T)^T$, and $\phi = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$. The log-likelihood function for model (2) is

$$l(\phi) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\beta})^T\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\beta}), \tag{4}$$

where $\boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma_0^2 + \sum_r^R \sum_{j\in O_r} \boldsymbol{K}_j^r\sigma_{rj}^2$. The corresponding penalized log-likelihood function with L_1 penalty is,

$$l_{\mathrm{p}}(\boldsymbol{\phi}) = l(\boldsymbol{\phi}) - \lambda_1|\boldsymbol{\omega_1\theta}|_1 - \lambda_2|\boldsymbol{\omega_2\beta}|_1, \tag{5}$$

where λ_1 and λ_2 are non-negative regularization parameters for random effects and fixed effects, respectively. $|A|_1$ is the l_1 norm of A. $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^T, \boldsymbol{\omega}_2^T)^T$ is adaptive weights, typically $\boldsymbol{\omega} = 1/|\tilde{\boldsymbol{\phi}}|$, with $\tilde{\boldsymbol{\phi}}$ denoting an initial $\sqrt{n}$ consistent estimator of ϕ (e.g. the maximum likelihood estimators). For both fixed and random effects, if we do not wish to perform variable selection for a particular parameter, we set the corresponding weights to be zero (e.g. if we intend to include all demographic variables for prediction, the adaptive weights that correspond to these demographic variables are set to be zero). Maximizing $l_{\mathrm{p}}(\boldsymbol{\phi})$ will enable variable selection and parameter estimation simultaneously, as the effects of less important factors will be shrunk to zeros under the L_1-penalty. The regularization parameters $(\lambda_i\boldsymbol{\omega}_i, \forall i \in (1, 2))$ are allowed to vary with the corresponding effect sizes, which is similar to the idea of adaptive lasso (Zou, 2006). It is worth noting that our selection scheme is very flexible as it allows for incorporating prior information into the adaptive weights, and thus the prior belief with regard to the importance of each predictor can be considered.

Maximizing Equation (5) can be computationally demanding, and thus we follow the procedure used in Lin *et al.* (2013a) and adopt a two-stage model selection procedure for the penalized LMM. The details of our proposed algorithm are depicted in Algorithm 1. We first maximize the penalized restricted log-likelihood function to select the random effects where a Newton–Raphson type algorithm is used, and then maximize the penalized log-likelihood function to select fixed effects. The penalized restricted log-likelihood function is given by,

$$Q_{\mathrm{R}}(\boldsymbol{\theta}) = l_{\mathrm{R}}(\boldsymbol{\theta}) - \lambda_1|\boldsymbol{\omega_1\theta}|_1, \tag{6}$$

where $l_{\mathrm{R}}(\boldsymbol{\theta})$ is the restricted log-likelihood function. Note that $\boldsymbol{\theta}$ is a $(\sum_r^R |O_r| + 1) \times 1$ dimensional vector, where $|O_r|$ represents the cardinality of set O_r (i.e. the total number of random effects associated with the rth region). We do not wish to select σ_0^2, and thus set $\omega_{10} = 0$. We apply an iterative procedure to estimate parameters $\boldsymbol{\theta}$. As the objective function (6) is non-differentiable at the origin, similar to (Fan and Li, 2001, 2012; Lin *et al.*, 2013b), we use a quadratic function to locally approximate the penalty function. Given the estimates are close to the maximizer of function (6), if the jth variable at iteration step s (denoted by θ_j^s) is very close to zero (i.e. $|\theta_j^s| < \delta$), we set the corresponding penalty function to be zero (i.e. $|\theta_j| = 0$). Otherwise, we use a local quadratic function to approximate the penalty function as,

$$|\theta_j| \approx \frac{1}{2}|\theta_j^s| + \frac{1}{2}\frac{(\theta_j)^2}{|\theta_j^s|}. \tag{7}$$

At iteration step $s+1$, we set $\theta_j^{s+1}=0$ if $|\theta_j^s|<\delta$. The rest parameters of $\boldsymbol{\theta}^{s+1}$ are estimated by solving Equation (8), where a Newton–Raphson algorithm is used.

$$\boldsymbol{\theta}^{s+1}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left\{l_{\mathrm{R}}(\boldsymbol{\theta})-\lambda_1\sum_{j=1}\omega_{1j}\frac{(\theta_j)^2}{2|\theta_j^s|}I(|\theta_j^s|\geq\delta)\right\}. \tag{8}$$

For the $p\times 1$ dimensional fixed effects $\boldsymbol{\beta}$, the penalized log-likelihood function when the covariance matrix of random effects is known, is given by,

$$Q_f(\boldsymbol{\beta})=-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\beta})-\lambda_2\sum_l\omega_{2l}|\beta_l|. \tag{9}$$

Maximizing function (9) is similar to solving an adaptive lasso problem (Zou, 2006), and thus efficient algorithms such as the least angel regression can be used to obtain parameter estimates (Efron *et al.*, 2004).

For the selection of tuning parameters, we apply the BIC-type criteria given by

$$\begin{cases}\mathrm{BIC}_{\lambda_1}=-2l_{\mathrm{R}}(\hat{\boldsymbol{\theta}})+\log(N)df_{\lambda_1}\\ \mathrm{BIC}_{\lambda_2}=-2l_f(\hat{\boldsymbol{\beta}})+\log(N)df_{\lambda_2}\end{cases}$$

for the random and fixed effects, respectively. Note that $l_{\mathrm{R}}(\hat{\boldsymbol{\theta}})$ is $l_{\mathrm{R}}(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$ and $l_f(\hat{\boldsymbol{\beta}})$ is $l_f(\boldsymbol{\beta})$ evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$. df_{λ_1} and df_{λ_2} are the number of non-zero elements in $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, respectively.

Algorithm 1. Two-stage procedure for estimating parameters

1: The penalized log-likelihood function is approximately maximized by a two-stage procedure, where $Q_{\mathrm{R}}(\boldsymbol{\theta})=l_{\mathrm{R}}(\boldsymbol{\theta})-\lambda_1|\boldsymbol{\omega_1\theta}|_1$ (stage 1) and $Q_f(\boldsymbol{\beta})=-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\beta})-\lambda_2|\boldsymbol{\omega_2\beta}|_1$ (stage 2) are maximized to estimate parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively.
2: Maximize $Q_{\mathrm{R}}(\boldsymbol{\theta})$ to estimate parameters $\boldsymbol{\theta}$.
  2.1: Get an initial estimates $\boldsymbol{\theta}^0$, where $l_{\mathrm{R}}(\boldsymbol{\theta})$ is maximized at $\boldsymbol{\theta}=\boldsymbol{\theta}^0$.
  2.2: Set the adaptive weights $\boldsymbol{\omega}_1$ for $\boldsymbol{\theta}$ with $\omega_{10}=0$.
  2.3: **for** ($t=1$ *to a sequence of* n_{λ_1} *tuning parameters of* λ_1) **do**
    while ($s\leq$ *max iteration* & $\boldsymbol{\theta}^s$ *at* $\lambda_{1,t}$ *does not converge*) **do**
      for ($j=1$ *to* $\sum^R|O_r|$) **do**
        if ($|\theta_j^s|<\delta$)' **then**
          Set the penalty term $|\theta_j|$ as 0 and $\theta_j^{s+1}=0$.
        else
          Use Equation (7) to approximate $|\theta_j|$.
        end if
      end for
      Use Newton–Raphson algorithm to maximize Equation (8),
      where L_1 penalty is approximated.
    end while
    Set $\hat{\boldsymbol{\theta}}_{\lambda_{1,t}}$ equal to $\boldsymbol{\theta}^s$ at convergence.
    Calculate $\mathrm{BIC}_{\lambda_{1,t}}=-2l_{\mathrm{R}}(\hat{\boldsymbol{\theta}}_{\lambda_{1,t}})+\log(N)df_{\lambda_{1,t}}$
  end for
  2.4: Choose λ_1 as $\lambda_1=\mathrm{argmin}_{\lambda_{1,t}}\mathrm{BIC}_{\lambda_{1,t}}$.
  2.5: Get the parameter estimates $\hat{\boldsymbol{\theta}}$ at λ_1.
3: Maximize $Q_f(\boldsymbol{\beta})$ to estimate parameters $\boldsymbol{\beta}$.
  3.1: Get the estimated variance–covariance matrix $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$.
  3.2: Solve an adaptive lasso problem with $\boldsymbol{Y}^*=\boldsymbol{A}^{-1}\boldsymbol{Y}$, $\boldsymbol{X}^*=\boldsymbol{A}^{-1}\boldsymbol{X}$ and $\boldsymbol{A}^T\boldsymbol{A}=\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$.
  3.3: Choose the tuning parameter λ_2 such that the *BIC* defined as $BIC_{\lambda_{2,t}}=-2l_f(\hat{\boldsymbol{\beta}}_{\lambda_{2,t}})+\log(N)df_{\lambda_{2,t}}$ is minimized.
  3.4: Get the parameter estimates $\hat{\boldsymbol{\beta}}$ at λ_2.
4: Build predictive model with the estimated parameters.

2.3 Theoretical results

The similarity matrices $(\boldsymbol{K}_j^r)$ are usually dense matrices for high-dimensional multi-layer omics data, and thus the variance–covariance matrix of $\boldsymbol{Y}$ $(\mathrm{var}(\boldsymbol{Y})=\sigma_0^2\boldsymbol{I}_n+\sum_r^R\sum_{j\in O_r}\boldsymbol{K}_j^r\sigma_{rj}^2)$ is also a dense matrix. The outcome vector $\boldsymbol{Y}$ is a single observation from a multivariate normal distribution, and thus the standard asymptotic theory for LMMs used in longitudinal and clustered data analyses is not directly applicable.

Denote the true values of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_0=(\boldsymbol{\theta}_{10}^T,\boldsymbol{\theta}_{20}^T)^T$, where $\boldsymbol{\theta}_{10}\neq\boldsymbol{0}$ is a vector that includes σ_0^2 and non-zero components of $\boldsymbol{\theta}_{\mathrm{omics}}$ and $\boldsymbol{\theta}_{20}=\boldsymbol{0}$ is a vector of the remaining components of $\boldsymbol{\theta}_{\mathrm{omics}}$. Similarly, we decompose $\boldsymbol{\theta}$ into $\boldsymbol{\theta}=(\boldsymbol{\theta}_1^T,\boldsymbol{\theta}_2^T)^T$, with $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ corresponding to $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{20}$. Let $N_{k_{10}}$ denotes the total number of random effects that is not zero, and $N_{k_{20}}$ the total number of random effects that is zero. Therefore, $\boldsymbol{\theta}_1$ is a $(N_{k_{10}}+1)\times 1$ dimensional vector and $\boldsymbol{\theta}_2$ is a $N_{k_{20}}\times 1$ dimensional vector. The total number of random effects (denoted by N_k) is $N_{k_{10}}+N_{k_{20}}$. Let the true value of $\boldsymbol{\beta}_0=(\boldsymbol{\beta}_{10}^T,\boldsymbol{\beta}_{20}^T)^T$, where $\boldsymbol{\beta}_{10}\neq\boldsymbol{0}$ is a $p_1\times 1$ vector whose components are not zero, and $\boldsymbol{\beta}_{20}=\boldsymbol{0}$ are the remaining zero components of fixed effects.

For our model, we use the general theory from Cressie and Lahiri (1993) and the results from Fan and Li (2001) to establish the asymptotic behavior of our estimators. The assumptions of our model are in Supplementary Appendix A.1. The assumptions (S.1)–(S.3) are similar to those used in Cressie and Lahiri (1993). Together with assumption (S.4), they yield a central limit theorem for $l'_{\mathrm{R}}(\boldsymbol{\theta})$ and convergence in probability of $l''_{\mathrm{R}}(\boldsymbol{\theta})$.

LEMMA 1. Under assumptions (S.1)–(S.4) in Supplementary Appendix A.1, for any $\boldsymbol{\theta}\in\Theta$, as $n\to\infty$, we have

$$n^{-1/2}l'_{\mathrm{R}}(\boldsymbol{\theta})\to_d N(\boldsymbol{0},\boldsymbol{J}(\boldsymbol{\theta}))\ \text{ and }\ n^{-1}l''_{\mathrm{R}}(\boldsymbol{\theta})\to_{\mathrm{p}}-\boldsymbol{J}(\boldsymbol{\theta})$$

THEOREM 1. Under assumptions (S.1)–(S.5) in Supplementary Appendix A.1, we have

a. *If* $\lambda_1/\sqrt{n}\to 0$, *then there exists a* $\sqrt{n}$*-consistent local maximizer of* $Q_{\mathrm{R}}(\boldsymbol{\theta})$.

b. *If* $\lambda_1\to\infty$, *then for any* $\hat{\boldsymbol{\theta}}_1$ *satisfying* $||\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}||\leq Mn^{-1/2}$ *and* $M>0$, $P(\hat{\boldsymbol{\theta}}_2=\boldsymbol{0})\to 1$.

c. *As* $\lambda_1\to\infty$ *and* $\lambda_1/\sqrt{n}\to 0$, *we have* $\sqrt{n}J(\boldsymbol{\theta}_{10})[\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}+\frac{\lambda_1}{n}J(\boldsymbol{\theta}_{10})^{-1}h(\boldsymbol{\theta}_{10})]\to_d N(0,J(\boldsymbol{\theta}_{10}))$, *where* $J(\boldsymbol{\theta}_{10})$ *is the upper-left sub-matrix of* $J(\boldsymbol{\theta})$ *and* $h(\boldsymbol{\theta}_{10})=(\omega_{10}\mathrm{sgn}(\theta_{10}^0),\omega_{11}\mathrm{sgn}(\theta_{10}^1),\ldots,\omega_{1N_{k_{10}}}\mathrm{sgn}(\theta_{10}^{N_{k_{10}}}))$ *with* θ_{10}^j *being the* j*th element in vector* $\boldsymbol{\theta}_{10}$ *and* $\omega_{10}=0$.

THEOREM 2. Under assumptions (S.1)–(S.5) given in Supplementary Appendix A.1, we have

a. *If* $\lambda_2/\sqrt{n}\to 0$, *then there exists a* $\sqrt{n}$*-consistent local maximizer of* $Q_f(\boldsymbol{\beta})$.

b. *If* $\lambda_2\to\infty$, *then for any* $\hat{\boldsymbol{\beta}}_1$ *satisfying* $||\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}||\leq Mn^{-1/2}$ *and* $M>0$, $P(\hat{\boldsymbol{\beta}}_2=\boldsymbol{0})\to 1$.

c. *As* $\lambda_2\to\infty$ *and* $\lambda_2/\sqrt{n}\to 0$, $\sqrt{n}J(\boldsymbol{\beta}_{10})[\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+\frac{\lambda_2}{n}J(\boldsymbol{\beta}_{10})^{-1}h(\boldsymbol{\beta}_{10})]\to_d N(0,J(\boldsymbol{\beta}_{10}))$, *where* $J(\boldsymbol{\beta}_{10})$ *is the upper-left sub-matrix of* $J(\boldsymbol{\beta})$ *and* $h(\boldsymbol{\beta}_{10})=(\omega_{21}\mathrm{sgn}(\beta_{10}^1),\omega_{22}\mathrm{sgn}(\beta_{10}^2),\ldots,\omega_{2p_1}\mathrm{sgn}(\beta_{10}^{p_1}))$ *with* β_{10}^j *being the* j*th element in vector* $\boldsymbol{\beta}_{10}$.

Theorem 1 says that (i) $\hat{\boldsymbol{\theta}}_{\lambda_1}$ is a $\sqrt{n}$-consistent estimator, (ii) the true model can be identified and (iii) $\hat{\boldsymbol{\theta}}_1$ is asymptotically normal. Theorem 2 suggests that given $\hat{\boldsymbol{\theta}}$, we have (i) $\hat{\boldsymbol{\beta}}_{\lambda_2}$ is a $\sqrt{n}$-consistent estimator, (ii) the true model can be identified with probability tending to 1 and (iii) $\hat{\boldsymbol{\beta}}_1$ is asymptotically normal. From these theorems, asymptotically our model can correctly identify predictive regions

and their estimated effects are normally distributed. The details of the proof are shown in Supplementary Appendix A.2.

2.4 Kernel principle component analysis for LMM

MKpLMM can accommodate potential correlations among different types of effects (e.g. the potential correlation between genotype and methylation effects) by constructing kernel functions on multiple layers of omics data, where the potential correlations are captured by the kernel matrix. Therefore, in practice, we are not particularly concerned about the correlation. Nevertheless, here we still present an alternative approach to model the potential correlations among different effects.

We consider the general model specified in Equation (2). For simplicity and without loss of generality, we consider only two types of effects (e.g. genotype and methylation effects) and treat them as random effects. We assume there are correlations between their effects. For simplicity, we only consider one region, but the same procedure applies when the number of regions is large. The model can be simplified as,

$$\boldsymbol{Y} = \boldsymbol{X_d\beta_d} + \sum_{r}^{R} \boldsymbol{O}^r + \epsilon = \boldsymbol{X_d\beta_d} + \boldsymbol{o}_g + \boldsymbol{o}_m + \epsilon, \tag{10}$$

where $\boldsymbol{o}_g \perp\!\!\!\perp \boldsymbol{o}_m$, $\boldsymbol{o}_g \sim N(0, \boldsymbol{K}_g\sigma_g^2)$, $\boldsymbol{o}_m \sim N(0, \boldsymbol{K}_m\sigma_m^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n\sigma_0^2)$. Model (10) is equivalent to

$$\boldsymbol{Y} = \boldsymbol{X_d\beta_d} + h_1(\boldsymbol{G}) + h_2(\boldsymbol{M}) + \boldsymbol{\varepsilon}, \tag{11}$$

where $\boldsymbol{G}$ and $\boldsymbol{M}$ are genotypes and methylation levels in the region. The functions h_1 and h_2 are from function spaces generated by kernel functions $\boldsymbol{K}_g$ and $\boldsymbol{K}_m$, respectively.

Following the same idea in Zhao *et al.* (2018), we use KPCA, a nonlinear version of PCA to transform h_1 and h_2 into linear space, where projections can be made. Consider the eigen decomposition of the kernel matrix, $\boldsymbol{K}_i = \boldsymbol{V}_i\Lambda_i\boldsymbol{V}_i^T, i \in \{g, m\}$, where $\boldsymbol{V}_i$ is the eigen vector, and $\Lambda_i = \text{diag}(\lambda_{i,1} \geq \lambda_{i,2}, \ldots, \geq \lambda_{i,b_i})$ are associated positive eigen values. Let $\boldsymbol{Z}_i = \boldsymbol{V}_i\Lambda_i^{1/2}$, for $i \in \{g, m\}$. Model (11) can be written as

$$\boldsymbol{Y} = \boldsymbol{X_d\beta_d} + \boldsymbol{Z}_g\boldsymbol{\beta}_g + \boldsymbol{Z}_m\boldsymbol{\beta}_m + \boldsymbol{\varepsilon}, \tag{12}$$

where $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n\sigma_0^2)$, $\boldsymbol{\beta}_g \sim N(0, \sigma_g^2)$, $\boldsymbol{\beta}_m \sim N(0, \sigma_m^2)$ and $\boldsymbol{\beta}_g \perp\!\!\!\perp \boldsymbol{\beta}_m$. Re-parameterizing the model, we project all layers of omics data onto one level (e.g. project the methylation data on to the genotype data) and construct the model as,

$$\boldsymbol{Y} = \boldsymbol{X_d\beta_d} + \boldsymbol{Z}_g\boldsymbol{\gamma}_g + \boldsymbol{Z}_m^*\boldsymbol{\gamma}_m + \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{Z}_m^* = (\boldsymbol{I} - \boldsymbol{P}_g)\boldsymbol{Z}_m$ with $\boldsymbol{P}_g = \boldsymbol{Z}_m(\boldsymbol{Z}_m^T\boldsymbol{Z}_m)^{-1}\boldsymbol{Z}_m^T$. Clearly, $\boldsymbol{Z}_m^*$ lays in the space that is orthogonal to the space of $\boldsymbol{Z}_g$. The effects $\boldsymbol{\gamma}_g$ and $\boldsymbol{\gamma}_m$ can be considered independent. We can reconstruct kernel functions based on $\boldsymbol{Z}_g$ and $\boldsymbol{Z}_m^*$, and build predictive models using the established theorems presented in this work.

3 Results

3.1 Simulation studies

In all simulation studies described below, we considered three types of omics data, including gene expression, methylation and genomic data. To adequately evaluate our method, the simulated datasets should represent the realistic correlations between features of the same type (e.g. the co-expression levels of genes in a pathway) and across data types (e.g. methylation in the promoter regions represses the expression of a gene). Therefore, we use the InterSIM software, which simulates multiple interrelated data types with realistic intra/inter-relationships based on the TCGA ovarian cancer study, to generate gene expression and methylation data (Chalise *et al.*, 2016). Since InterSIM software does not simulate genomic data, to mimic the real human genome, we first extract all single nucleotide variants (SNVs) from the whole-genome sequencing data from the 1000 Genome Project (The 1000 Genomes Project Consortium, 2015) and then select SNVs located in the genomic regions with simulated gene expression and methylation data. We exclude genes on which there are no SNVs. The details of genes used in the simulation studies are summarized in Supplementary Table S1.

We map the SNVs and CpG sites into gene regions and simulate quantitative phenotypes based on causal genes. For all the simulations described below, we randomly choose four genes to be causal and simulate the phenotypes under various disease models. To account for the heterogeneous effects from the causal genes, the effect sizes in each gene are set in the ratios of $1 : 1.1 : 1.1^2 : 1.1^3$. Specifically, we simulate the disease outcomes under the general formula as

$$\begin{aligned} Y_i = \sum_{k}^{4} E_k\beta_k^e + \sum_{k}^{4}\sum_{j}^{n_k^g} G_{kj}\beta_{kj}^g I(P_{kj} = 1) + \sum_{k}^{4}\sum_{l}^{n_k^m} M_{kl}\beta_{kl}^m \\ + \sum_{k}^{4}\sum_{j}^{n_k^g}\sum_{l}^{n_l^m} G_{kj}M_{kl}\beta_{kjl}^{gm} I(P_{kj} = 1) + \epsilon_i, \end{aligned}$$

where E_k, G_{kj} and M_{kl} respectively represent gene expression level for gene k, genotype at the jth location of gene k and the methylation level at the lth CpG site of gene k, and β_k^e, β_{kj}^g and β_{kl}^m are their corresponding effects. β_{kjl}^{gm} represents the interaction effects between the genotype at the jth location and the methylation level at the lth CpG site of gene k. n_k^g and n_k^m are the number of SNVs and the number of CpG sites for gene k, respectively. We set $P_{kj} \sim Ber(0.25)$, and thus 25% of SNVs located on the causal genes are set causal. The details of disease models and effect sizes are summarized in Supplementary Table S2.

In the first set of simulations, we evaluate the impact of the number of noise genes on the performance of the method by gradually increasing the number of noise genes from 21 to 96 (i.e. the total number of genes changes from 25 to 100), where we assume gene expression levels, methylation levels and SNVs from each causal gene all contribute to disease risk (S7 in Supplementary Table S2). In the second set of simulations, we evaluate the performance of our method under different disease models. Specifically, we considered eight disease models including (i) only one type of omics data contributes to disease risk (S1–S3 in Supplementary Table S2), (ii) multiple types of omics data independently contribute to disease risk (S4–S7 in Supplementary Table S2) and (iii) multiple types of omics data jointly contribute to disease risk (S8 in Supplementary Table S2).

For each setting, we generate 500 Monte Carlo replicates. The sample size is set to be 1000 with 500 samples served as training samples. We build prediction models based on the training samples and evaluate the performance based on the testing samples. We use Pearson correlations and mean square errors (MSEs) to measure the prediction accuracy. We compare the performance of MKpLMM with OmicKrig (Wheeler *et al.*, 2014), a commonly used method for prediction analysis on multi-omics data. We further present the results where single-layer omics data is used for prediction. For the OmicKrig method (Wheeler *et al.*, 2014), we use their default settings with the default kernel functions. For MKpLMM, we use a gene-based approach (i.e. the genomic region is defined by the gene), and treat the gene expression levels as fixed effects. We grouped SNVs and methylation levels according to the gene, and treat their effects as random effects. We use the linear kernel function for both SNVs and methylation data to calculate region-wise genomic and epigenomic similarities, and use Hadamard product between them to capture the interactions among SNVs and methylation levels. We further calculate the probability of correctly selecting predictive genes for the MKpLMM method.

Figure 1 summarizes the Pearson correlations from the first set of simulations, and the MSEs are presented in Supplementary Figure S1. As expected, the performance of both methods increases as the effect sizes increase. For all the situations considered, our method has better performance than OmicKrig (Wheeler *et al.*, 2014), which indicates excluding noise regions from the analysis can improve the prediction accuracy. We also compare prediction performances of

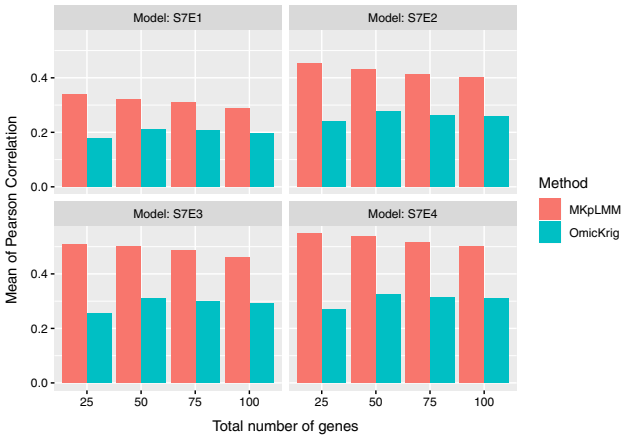

Fig. 1. The effects of the number of noise regions

Table 1. The probability of selecting the true causal genes under different number of noise genes

	Sensitivity (specificity)			
	No. gene = 25	No. gene = 50	No. gene = 75	No. gene = 100
S7E1	0.762 (0.898)	0.690 (0.939)	0.629 (0.957)	0.611 (0.964)
S7E2	0.919 (0.863)	0.863 (0.922)	0.808 (0.946)	0.747 (0.955)
S7E3	0.959 (0.822)	0.927 (0.904)	0.874 (0.940)	0.857 (0.945)
S7E4	0.962 (0.822)	0.940 (0.906)	0.902 (0.932)	0.842 (0.946)

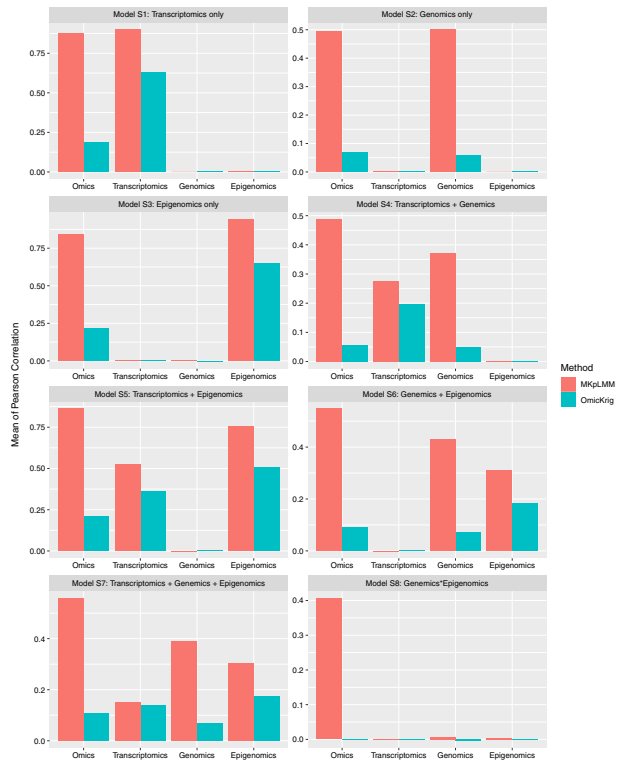

Fig. 2. The performance under different disease models

each method under the settings where (i) all omics data is used for prediction, (ii) only gene expression data is used, (iii) only SNV data is used and (iv) only methylation data is used. The results are shown in Supplementary Figure S2. For both OmicKrig and MKpLMM, including all available information (i.e. using mulit-omics data) can substantially improve prediction accuracy. This suggests that jointly analyzing all omics data can benefit prediction analyses when all omics data contribute to disease risk. We further summarize the probability of correctly selecting predictive genes for our method. As shown in Table 1, although the sensitivity and specificity can be affected by the total number of noise genes and the effect sizes, our method can achieve a sensitivity of 83% and a specificity of 92% on average. All specificities are above 82% and all sensitivities are above 61% among all the situations considered, indicating our method has the capacity of correctly selecting predictive genes and teasing out the effects of noise genes. We consider this important, as in reality many of the measured genomic regions do have any predictive effects.

Figure 2 summarizes the mean of Pearson correlations of the proposed method under different disease models where the effect sizes are set under E4 (Supplementary Table S2) and the total number of genes equals to 25. The rest results (i.e. effect sizes are under scenario E1–E3 of Supplementary Table S2 and the total number of genes are 25, 50, 75, 100) are shown in Supplementary Figure S3. The MSEs are summarized in Supplementary Figure S4. As indicated from these figures, our method performs robust against different disease models. When only one level of omics data contributes to disease risk (i.e. Supplementary Models S1–S3), our method performs similar to the one where only the relevant omics data that contributes to disease risk is used. For example, under Supplementary Model S1 where only gene expression data contributes to disease risk, our method performs similar to the one where only gene expression data is used whereas the performance of OmicKrig drops when all layers of multi-omics data is used. When multiple layers of omics data marginally contribute to disease risk (i.e. Supplementary Models S4–S7), our method performs better than the ones where only single-layer omics data is used. When multiple layers of omics data affect disease risk through interactions (i.e. Supplementary Model S8), our method outperforms the methods that only use single-layer omic data. Moreover, our method also performs substantially better than OmicKrig under this setting, as OmicKrig assumes that each layer of omic data contributes independently to disease risk and thus fails to capture the interaction effects. We consider the robustness against disease models important, as in practice the underlying mechanisms of diseases are usually unknown in advance. Therefore, a method that can adaptive choose kernel functions and predictive regions to capture various types of effects for risk prediction analyses is practically useful. To assess the selection performance, we further calculate the sensitivity and specificity for our method. Although the sensitivity and specificity of the selection can be affected by the number of noise genes, the effect sizes and the disease model, our method in general achieves relatively high specificity for all disease models and its sensitivity tends to be high when there is no interactions (Table 2 and Supplementary Table S3). On average, our method achieves a sensitivity of 84% and a specificity of 91% among all the models considered.

3.2 The analysis of Alzheimer's disease dataset

We analyzed the whole genome sequencing and gene expression data from ADNI using both the proposed method and OmicKrig with the default settings (Wheeler *et al.*, 2014). ADNI, including ADNI 1, ADNI GO and ADNI 2, is a longitudinal study that can be used to assess the effects of biomarkers at various levels on Alzheimer's Disease (AD). Study participants were followed and assessed over time to investigate the pathology of AD. DNA samples from study subjects in ADNI 2, including newly recruited subjects and ADNI 1/GO continuing participants, were obtained and analyzed using Illumina's non-CLIA whole genome sequencing. RNA expression data were collected from subjects in ADNI 2 at baseline for newly recruited subjects and 1st ADNI 2 visit for ADNO 1/GO continuing subjects and then yearly. Imaging data (e.g. MR imaging and PET imaging) were collected at each visit. For our analyses, we focus on baseline data, and we are interested in predicting PET-imaging outcomes (i.e. outcomes from FDG and AV45 scans) using both DNA sequencing data and gene expression data. The sample

sizes for genomic data, gene expression and the imaging outcomes are summarized in Supplementary Figure S5.

We annotated the genetic variants based on GRch38 assembly, and included a total of 96 genes that have been previously reported to be associated with AD. The complete list of the genes included in the analyses is summarized in Supplementary Table S4. After excluding the SNVs without any variations, 119 144 SNVs are included in the final analyses. The minor allele frequencies for these SNVs are presented in Supplementary Figure S6. For each selected gene, we assessed the correlations between gene expression levels and the outcome of interest, and also conducted a single-locus analyses for all SNVs included in the selected regions. The results are presented in Supplementary Figures S7 and S8.

For prediction analyses, we randomly selected 60 subjects to serve as the testing samples and used the remaining samples to build predictive models. We evaluated the prediction accuracy based on the testing samples using Pearson correlations and MSEs. To avoid the chance findings, we repeat this process 200 times. The prediction accuracy for AV45 and FDG are shown in Figure 3. For both AV45 and FDG, the Pearson correlations of the MKpLMM are higher and MSEs of MKpLMM are smaller than the OmicKrig method, suggesting MKpLMM achieves better prediction accuracy than the OmicKrig method. This indicates that excluding noise genes from the prediction can improve prediction accuracy. Comparing the prediction models built with omics data and the ones built with single level data, the FDG and AV45 can be mainly predicted by the genomic data, which is consistent with the analysis results shown in single gene analyses (Supplementary Figs S7 and S8). We further calculated the probability of each gene being selected by MKpLMM for AV45 and FDG. The MKpLMM achieves robust performance with regard to the variable selection (Supplementary Table S4). More than 36% of the genes have never been selected for both AV45 and FDG, and all the other genes except *APOE* have been selected <10%. The *APOE* gene on chromosome 19, a well-known risk predictor for AD, has been selected 98% for AV45 and 99% for FDG. The selection result is also consistent with the results from single locus analyses (Supplementary Figs S7 and S8), as most of the significant signals come from *APOE*.

Table 2. Probability of selecting true causal genes under different disease models

	Sensitivity (specificity)			
	No. gene = 25	No. gene = 50	No. gene = 75	No. gene = 100
S1E4	1.000 (0.990)	1.000 (0.995)	1.000 (0.991)	1.000 (0.986)
S2E4	0.990 (0.939)	0.987 (0.969)	0.965 (0.968)	0.921 (0.979)
S3E4	0.718 (0.588)	0.654 (0.787)	0.622 (0.868)	0.580 (0.892)
S4E4	1.000 (0.923)	0.997 (0.917)	0.986 (0.953)	0.974 (0.958)
S5E4	1.000 (0.670)	1.000 (0.818)	1.000 (0.871)	1.000 (0.900)
S6E4	0.930 (0.847)	0.901 (0.913)	0.855 (0.936)	0.804 (0.945)
S7E4	0.962 (0.822)	0.940 (0.906)	0.902 (0.932)	0.842 (0.946)
S8E4	0.972 (0.924)	0.672 (0.818)	0.575 (0.869)	0.469 (0.902)

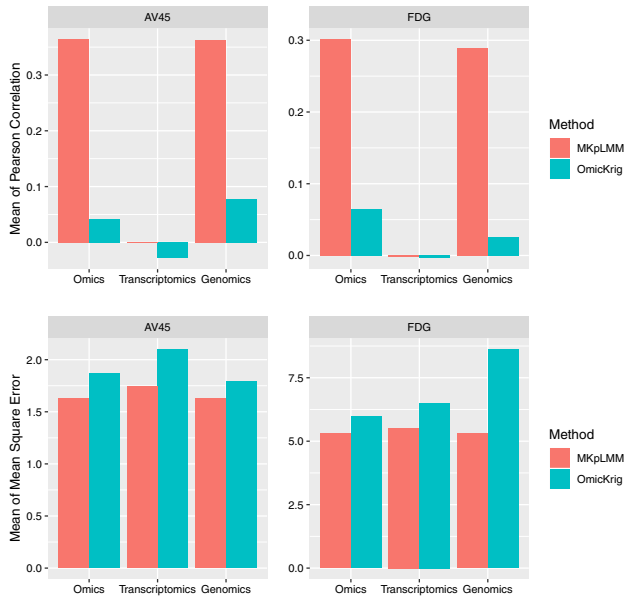

Fig. 3. The prediction accuracy for AV45 and FDG

3.3 The analysis of chronic lymphocytic leukemia dataset

We further applied MKpLMM to a study of CLL, which measured the *ex-vivo* drug responses, somatic mutation status, transcription and methylation profiles. The CLL study is designed to investigate the determinants of drug responses. The study recruited 246 subjects with 200 being CLL patients. These patients were profiled with 64 drugs in series of 5 concentrations. We are interested in using multi-omics data to predict drug responses among these 200 CLL patients. We constructed the drug response variable (i.e. drug viability) based on the procedures recommended by Dietrich *et al.* (2017). For each drug, we aim to predict both the average viability values across all 5 concentrations and across the lowest 2 concentrations. For genetic data, we followed the same procedure by Dietrich *et al.* (2017), and in total we included 11 genes for somatic mutations. For both methylation and gene expression data, we grouped them according to cancer pathways listed in the KEGG database and in total 95 270 CpG sites and 4428 gene expression levels are included in the analyses. The sample sizes for each omics data and drug responses are shown in Supplementary Figure S9, and in total we included 102 patients that have complete data for this analysis.

We assessed the association between somatic mutation status and the drug viabilities for each drug, and the results are shown in Supplementary Table S5. We further assessed the association between gene expression levels of selected genes and the outcomes of interest (i.e. the average of viabilities across 5 concentrations and under the lowest 2 concentrations), and the results are shown in Supplementary Figures S10 and S11. Finally, we assessed the association of methylation levels at each CpG site with the drug viabilities, and the results are shown in Supplementary Figures S12 and S13. For each of the prediction analyses, we randomly selected 90 subjects to serve as the training samples and used the remaining 12 individuals as the testing samples. We built the predictive models using training samples, and calculated the Pearson correlations and MSEs based on the testing samples. Similar to the ADNI data analyses, we repeat this process 200 times for each drug response outcome to avoid the chance finding. The comparisons of prediction accuracies between MKpLMM and OmicKrig for the average viabilities across all 5 concentrations and across the lowest 2 concentrations for all 64 drugs are shown in Figure 4. While the multi-omics data have different capacities of predicting the drug responses as shown in Supplementary Figures S10–S13, our proposed MKpLMM tends to perform similar or better than OmicKrig. The mean of Pearson correlations (Fig. 4) of the MKpLMM is higher than that of OmicKrig (i.e. the differences is larger than zero) in most drugs, and the MSEs (Supplementary Fig. S14) in MKpLMM is usually smaller than those in OmicKrig (i.e. MSE ratios are <1). This indicates that MKpLMM has similar if not better performance than OmicKrig in predicting the drug responses.

4 Discussion

We have presented MKpLMM, a powerful and efficient method for prediction of complex traits from multi-layer omics data. Our method generalizes the state-of-the-art LMM-based methods used for prediction analyses with single-layer genomic data to multi-

omics data, and yields better prediction performance on several datasets. MKpLMM is flexible with the input data types and can accommodate various types of effects through multiple kernel functions. MKpLMM has only one tuning hyper-parameter that determines the sparseness of the model and is chosen to provide balance between the model complexity and robustness, allowing for the detection of important predictors. Through simulation studies, we demonstrate that MKpLMM can successfully identify predictors with both linear and nonlinear effects, and achieves higher prediction accuracy than the OmicKrig method. Moreover, we also show that multi-omics data has the capacity to improve prediction accuracy compared to single-layer data analyses if analyzed appropriately. When applied to both drug and ANDI datasets, the MKpLMM achieves better or similar performance to the other method, and consistently selects the predictive genes.

MKpLMM is a LMM-based method, and thus it shares the advantages of LMM-based methods used for prediction with single-layer genomic data. Other prediction methods, such as support vector machine, Bayesian networks, penalized partial least square and deep learning are potential candidates for risk prediction analyses. Nevertheless, MKpLMM has several features that make it appealing for modeling multi-layer omics data. First, similar to LMMs and their extensions, MKpLMM has the capacity to handle high-dimensional data. MKpLMM assumes that omics similarities lead to phenotypic similarity, and encodes the omics effects through a covariance matrix that scaled quadratically with sample size regardless of the original data dimension. The multi-omics data analysis is under the setting of $n \ll p$, and thus MKpLMM substantially reduces the data dimension. Second, MKpLMM has the natural advantages of handling heterogeneous data types. The effects of predictors from various omics-layers are encoded through the covariance matrices, and thus successfully transform the prediction problem from heterogeneous high-dimensional feature spaces to a more homogeneous sample space. This property renders them particularly suitable for modeling heterogeneous multi-omics data. Third, the parameters in MKpLMM (i.e. fixed and random effect estimates for each omics layer) can be selected and inferred analytically. In this work, we have established the corresponding theory for parameter selection and estimation for both fixed and random effects under common settings in genetic research. It is well recognized that choosing the subset of important variables can substantially improve omics data integration performance. Detailed reviews of existing variable selection/dimension reduction methods for multi-omics data can be found in Wu *et al.* (2019) and Meng *et al.* (2016). Unsupervised/semi-supervised methods, such as canonical correlation analysis (Gross and Tibshirani, 2015; Witten and Tibshirani, 2009) and matrix factorization (Yang and Michailidis, 2016; Zhang *et al.*, 2012; Zitnik and Zupan, 2015), usually construct their loss function based on the distances between the lower-dimensional projected multi-omic matrix and the original data matrix, and their main focus is to understand the inter-relationships among multi-omics data. Supervised methods, however, usually construct their loss function based on the distances between the predicted outcomes and the original outcomes, and thus have a natural advantage for predicting the phenotypes of interest (Jiang *et al.*, 2016; Zhu *et al.*, 2016). Our proposed KMpLMM can be viewed as a supervised learning problem with objective function of the form 'unpenalized loss function + penalty function'. Variable selection of our method is achieved through penalization, which is widely used in bioinformatic applications (Wu and Ma, 2015; Wu *et al.*, 2019). Our key contribution is to establish the analytical theory for parameter selection, especially for random effects selection, in LMMs under the setting of genetic studies. This has not been extensively studied in the existing literature (Lin *et al.*, 2013a) and empirical criteria are often used instead (Speed and Balding, 2014; Weissbrod *et al.*, 2016). The analytical framework for model selection is important, as it allows for efficiently and simultaneously infer a large number of model parameters including both fixed and random effects.

MKpLMM is a flexible framework that not only allows for incorporating prior knowledge about the types of effects through the choice of kernel functions, but also allows for the selection of appropriate kernel functions in a data-driven manner. For example, both linear kernel that captures the additive effects and polynomial kernel with degree 2 that is often used in genetic studies to capture pairwise interactions can be specified for each genomic region, and MKpLMM can select the best kernel functions for each region in a data-driven manner. The selection of both predictive genomic regions and the appropriate kernels makes MKpLMM robust to various underlying disease models. As shown in our simulation studies, MKpLMM has the capacity of capturing different types of effects from various layers of omics data. Its performance substantially outperforms OmicKrig, when the outcomes are only affected through interactions. While in this study we only consider the linear kernel and polynomial kernel with degree 2 for each genomic

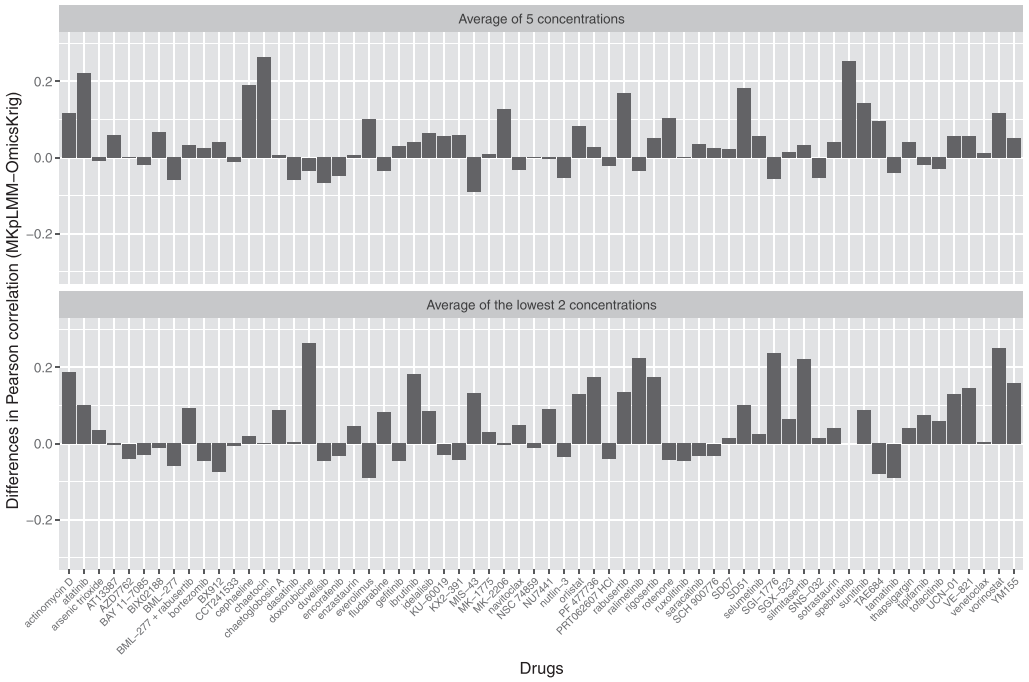

Fig. 4. The comparisons of prediction accuracies between MKpLMM and OmicKrig

region, there is a wide range of kernels in the existing literatures (e.g. RBF kernel for gene expression data, IBS kernel for genomic data and saturate pathway models for capturing nonlinear effects) and they can be incorporated into MKpLMM in a straightforward manner. Our method also allows for incorporating prior knowledge on the importance of each region, where the adaptive weights are specified to reflect domain knowledge. In the absence of prior knowledge, our method can use a data-driven manner to determine the relative importance of each region, where the adaptive weights are assigned as $1/\tilde{\phi}$ with $\tilde{\phi}$ being a $\sqrt{n}$ consistent estimator.

While MKpLMM is particularly useful for capturing interactions between variants located in the same region, distant interactions can also be exploited for improved prediction through the construction of pseudo-region encompassing these variants. Under the default settings of MKpLMM, we do not consider distant interactions because exhaustively evaluating all possible distant interactions can exponentially increase the number of parameters. This can substantially increase the computational burden, and may low the prediction performance due to the addition of a huge number of variances into the model. However, in practice, if a researcher has suspected potential interactions among distant regions, MKpLMM is capable of exploiting these interaction effects by manually constructing regions with these suspected variants harbored. Similarly, the default settings of MKpLMM only focus on capturing the interactions among different omics layers within the same region. However, MKpLMM is capable of capturing interactions among different layers of omics from different genomic regions by constructing a similarity matrix where the interactions between different layers of omics from different genomic regions are considered. This similarity matrix will be treated the same as the other region-similarity matrices, and MKpLMM can determine whether the distant interactions can improve prediction in a data-driven manner.

MKpLMM can be viewed as using a composite kernel with a weighted average of each single kernel, where each weight represents the effects from the corresponding data layer. The optimal weight depends on the true nature of omics effects that is usually unknown in advance, and thus MKpLMM adopts a data-adaptive approach to estimate and select them. The proposed model is equivalent to $E(\boldsymbol{Y}) = \boldsymbol{X\beta} + \sum_i^R [h_i(\boldsymbol{G}_i, \boldsymbol{E}_i, \boldsymbol{M}_i)]$, with h_i being function spaces generated by kernel functions K_i. The statistical inference for LMMs used in genomic research largely relies on the assumption that the effects from genomic variants are independently distributed. In this project, we use a similar assumption where we assume h_i is independent of each other. However, this assumption may not hold for multi-layer omics data. To accommodate the potential correlations, we also developed MKpLMM with KPCA. We use KPCA to transform the unknown functions h_i into linear functions and further project the multi-layer omics data onto a common space and an orthogonal space for each omics layer. This procedure is similar to the approach that includes one layer of omics data (e.g. genomic) as covariates and treat information from other layers (e.g. additional effects of methylation) as additional effects beyond the base layer on the outcome. While MKpLMM with KPCA can guarantee the theoretical ground for statistical inference, it may lose power due to the linearization. Therefore, in the prediction-based statistical learning, we recommend to use MKpLMM. However, if the main purpose of the study is to detect which genomic regions and which layers of omics data are associated with the outcomes, MKpLMM with KPCA is an alternative choice.

In our simulation studies, we grouped predictors according to gene annotations, and thus for each gene region we had only one variable for gene expression level and multiple variables for both methylation and genetic predictors. While in simulation studies we treat gene expression data as fixed effects and the others as random effects, we can also treat all layers of omics data as random effects (performance is shown in Supplementary Fig. S15 and Table S6). Although both strategies can result in prediction models with similar levels of performance, the computational complexity for a model with k fixed effects is much less than that of the model with k random effects. (The computational time as the number of random effects increases in shown in Supplementary Fig. S16.) Therefore, we recommend to treat the omics layer with only one predictor per region (e.g. gene expression level for each gene) for all k regions as k fixed effects.

In the illustrations with real datasets, we applied our methods to both the ADNI and the drug response datasets. As shown in Figures 3 and 4, the MKpLMM has better or similar prediction performance as compared to the existing method. Moreover, similar to simulation studies, multi-omics data can help to improve prediction accuracy if integrated appropriately. For the selection, our algorithm is in general consistent. For ADNI, the most commonly selected gene is the *APOE*, a well-known risk factor for AD.

The work introduced in this paper focuses on continuous outcomes that are normally distributed. The analysis of binary outcomes within the LMM framework can be challenging due to the intractable parameter inference. While existing studies have demonstrated that LMMs can achieve reasonable prediction performance when the binary outcomes are treated as if they were normally distributed (Speed and Balding, 2014; Weissbrod *et al.*, 2016), it would be interesting to investigate, within the framework of generalized LMM, other link functions (e.g. logit and log) for the prediction of outcomes with various distributions (e.g. binary and Poisson) for multi-omics data analysis.

A potential limitation of the proposed model is that we treat all the outcomes as if they had the same causes. Most of the common diseases (e.g. cancer and spectrum disorders) are heterogeneous in nature, and thus allowing different layers of omics data to contribute differently according to the underlying causes have the potential to substantially increase prediction accuracy and allow for the identification of subgroups of patients that are sensitive to various treatments. A natural way to incorporate disease heterogeneity for prediction modeling within the LMM framework is to relax the assumption on the distribution of effect sizes. Instead of using the common assumption that the variants within the genomic regions follow a normal distribution [i.e. $\beta \sim N(0, \sigma^2)$], the effect sizes can be assumed to follow a mixture of multivariate Gaussian with different means and a common covariance matrix [i.e. $\beta \sim \sum_g \pi_g N(\mu_g, D)$]. Similar theory and algorithms developed for MKpLMM can be adapted to infer parameters and be used for predictions, and this will be a future direction of our research. While our main focus of this study is to make accurate prediction using high-dimensional multi-layer omics data, MKpLMM especially MKpLMM with KPCA has the potential to be adapted for other tasks (e.g. testing for association with detect disease-associated genomic regions/pathways). This remains a potential avenue for our future research.

In summary, we have developed an MKpLMM for prediction analyses on high-dimensional multi-layer omics data. Through both simulation and real data applications, we have demonstrated that compared to single-layer data analyses, integrating multi-omics data using a data-driven approach to capture potential interactions among omics data can substantially improve prediction accuracy.

Acknowledgements

We wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. We wish to thank the reviewers for their valuable comments.

Funding

The Project was supported by the National Natural Science Foundation of China [Award No. 81502887], the Faculty Research Development Funds from the University of Auckland, the National Institute on Drug Abuse [Award No. R01DA043501] and the National Library of Medicine [Award No. R01LM012848].

Conflict of Interest: none declared.

References

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

Bersanelli,M. *et al.* (2016) Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinform.*, **17** (**Suppl. 2**), 15.

Boekel,J. *et al*. (2015) Multi-omic data analysis using galaxy. *Nat. Biotechnol*., **33**, 137–139.
Buil,A. *et al*. (2015) Gene–gene and gene–environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet*., **47**, 88–91.
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.
Chalise,P. *et al*. (2016) Intersim: simulation tool for multiple integrative 'omic datasets'. *Comput. Methods Programs Biomed*., **128**, 69–74.
Chen,J. and Zhang,S. (2016) Integrative analysis for identifying joint modular patterns of gene-expression and drug–response data. *Bioinformatics*, **32**, 1724–1732.
Cho,D.Y. and Przytycka,T.M. (2013) Dissecting cancer heterogeneity with a probabilistic genotype–phenotype model. *Nucleic Acids Res*., **41**, 8011–8020.
Cressie,N. and Lahiri,S.N. (1993) The asymptotic-distribution of REML estimators. *J. Multivariate Anal*., **45**, 217–233.
Dietrich,S. *et al*. (2017) Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest*., **128**, 427–445.
Efron,B. *et al*. (2004) Least angle regression. *Ann. Stat*., **32**, 407–451.
Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc*., **96**, 1348–1360.
Fan,Y. and Li,R. (2012) Variable selection in linear mixed effects models. *Ann. Stat*., **40**, 2043–2068.
Fisher,V.A. *et al*. (2018) Do changes in dna methylation mediate or interact with SNP variation? A pharmacoepigenetic analysis. *BMC Genet*., **19** (**Suppl. 1**), 70.
Gross,S.M. and Tibshirani,R. (2015) Collaborative regression. *Biostatistics*, **16**, 326–338.
Jiang,Y. *et al*. (2016) Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics*, **107**, 223–230.
Lin,B.Q. *et al*. (2013a) Fixed and random effects selection by REML and pathwise coordinate optimization. *J. Comput. Graph. Stat*., **22**, 341–355.
Lin,D. *et al*. (2013b) Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinform*., **14**, 245.
Lock,E.F. *et al*. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat*., **7**, 523–542.
Meng,C. *et al*. (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform*., **17**, 628–641.
Moore,J.H. and Williams,S.M. (2009) Epistasis and its implications for personal genetics. *Am. J. Hum. Genet*., **85**, 309–320.
Morris,J.S. and Baladandayuthapani,V. (2017) Statistical contributions to bioinformatics: design, modelling, structure learning and integration. *Stat. Model*., **17**, 245–289.
Ritchie,M.D. *et al*. (2015) Methods of integrating data to uncover genotype–-phenotype interactions. *Nat. Rev. Genet*., **16**, 85–97.
Saykin,A.J. *et al*. (2010) Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement*., **6**, 265–273.
Shen,R. *et al*. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
Speed,D. and Balding,D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res*., **24**, 1550–1557.
Speicher,N.K. and Pfeifer,N. (2015) Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, **31**, i268–275.
The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
VanRaden,P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci*., **91**, 4414–4423.
Wang,B. *et al*. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
Weissbrod,O. *et al*. (2016) Multikernel linear mixed models for complex phenotype prediction. *Genome Res*., **26**, 969–979.
Wheeler,H.E. *et al*. (2014) Poly-omic prediction of complex traits: OmicKriging. *Genet. Epidemiol*., **38**, 402–415.
Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol*., **8**, Article 28.
Wu,C. and Ma,S. (2015) A selective review of robust variable selection with applications in bioinformatics. *Brief Bioinform*., **16**, 873–883.
Wu,C. *et al*. (2019) A selective review of multi-level omics data integration using variable selection. *High Throughput*, **8**, pii: E4.
Yang,J. *et al*. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet*., **42**, 565–569.
Yang,J. *et al*. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*., **88**, 76–82.
Yang,Z. and Michailidis,G. (2016) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, **32**, 1–8.
Zeng,I.S.L. and Lumley,T. (2018) Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinform. Biol. Insights*, **12**, 117793221875929.
Zhang,S. *et al*. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*., **40**, 9379–9391.
Zhao,N. *et al*. (2018) Kernel machine methods for integrative analysis of genome-wide methylation and genotyping studies. *Genet. Epidemiol*., **42**, 156–167.
Zhu,R. *et al*. (2016) Integrating multidimensional omics data for cancer outcome. *Biostatistics*, **17**, 605–618.
Zitnik,M. and Zupan,B. (2015) Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell*., **37**, 41–53.
Zou,H. (2006) The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc*., **101**, 1418–1429.