Hindawi
Computational and Mathematical Methods in Medicine
Volume 2019, Article ID 2807470, 8 pages
https://doi.org/10.1155/2019/2807470

Research Article

Jackknife Model Averaging Prediction Methods for Complex Phenotypes with Gene Expression Levels by Integrating External Pathway Information

Xinghao Yu,[1] **Lishun Xiao**,[1] **Ping Zeng**,[1,2] **and Shuiping Huang**[1,2]

[1] *Department of Epidemiology and Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu 221004, China*
[2] *Center for Medical Statistics and Data Analysis, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu 221004, China*

Correspondence should be addressed to Ping Zeng; zpstat@xzhmu.edu.cn and Shuiping Huang; hsp@xzhmu.edu.cn

Received 13 January 2019; Accepted 20 March 2019; Published 8 April 2019

Academic Editor: Nadia A. Chuzhanova

Motivation. In the past few years many prediction approaches have been proposed and widely employed in high dimensional genetic data for disease risk evaluation. However, those approaches typically ignore in model fitting the important group structures that naturally exists in genetic data. *Methods*. In the present study, we applied a novel model-averaging approach, called jackknife model averaging prediction (JMAP), for high dimensional genetic risk prediction while incorporating pathway information into the model specification. JMAP selects the optimal weights across candidate models by minimizing a cross validation criterion in a jackknife way. Compared with previous approaches, one of the primary features of JMAP is to allow model weights to vary from 0 to 1 but without the limitation that the summation of weights is equal to one. We evaluated the performance of JMAP using extensive simulation studies and compared it with existing methods. We finally applied JMAP to four real cancer datasets that are publicly available from TCGA. *Results*. The simulations showed that compared with other existing approaches (e.g., gsslasso), JMAP performed best or is among the best methods across a range of scenarios. For example, among 14 out of 16 simulation settings with $\mathrm{PVE} = 0.3$, JMAP has an average of 0.075 higher prediction accuracy compared with gsslasso. We further found that in the simulation, the model weights for the true candidate models have much smaller chances to be zero compared with those for the null candidate models and are substantially greater in magnitude. In the real data application, JMAP also behaves comparably or better compared with the other methods for continuous phenotypes. For example, for the COAD, CRC, and PAAD datasets, the average gains of predictive accuracy of JMAP are 0.019, 0.064, and 0.052 compared with gsslasso. *Conclusion*. The proposed method JMAP is a novel model-averaging approach for high dimensional genetic risk prediction while incorporating external useful group structures into the model specification.

1. Introduction

Due to the rapid development of biotechnology [1–4], a large number of high-throughput and low-cost genetic datasets have been generated and provide a broad space to investigate the association between genetic markers and complex diseases/disorders [5–14]. The great success of association studies further promotes the risk prediction and evaluation for complex phenotypes by incorporating into genetic information (e.g., gene expressions or single nucleotide polymorphisms) [15–20]. Due to the high dimensional problem that the number of genetic markers is much larger than the sample size, one of the greatest challenges for genetic risk prediction is that it is difficult to apply traditional statistical methods in large scale molecular omics data. In the past few years, developing prediction methods that can efficiently model high dimensional genetic data has been an active area and attracted much research attention, and a series of novel prediction approaches have been proposed and widely employed for disease risk evaluation or gene expression imputation [21–27]. However, most of those approaches ignore in model fitting the important information of group structures or functional classifications that naturally exist in genetic data. For example, it is well known that genes can be grouped into pathways due to the shared biological function [28]. It has been shown that incorporating such useful group/functional information into model fitting can substantially boost statistical power in genetic association studies and can facilitate our understanding of the genetic architecture of disease variation by heritability partition [27, 29–36]. In genetic data, one of the widely-used group sources is the pathway information in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [37, 38], which integrates information on genomic, chemical, and system functions and groups genes with highly related sequences in terms of the sequence similarity of genes.

Besides being included in genetic association studies and heritability estimation, group/functional information is also recently integrated into genetic risk evaluation with large scale omics data, e.g., the protein-network-based method [39] and the combined-optimal-response-genes (CORGs) approach [40]. Additionally, the regularization methods (e.g., group Lasso) can perform a group selection and estimation by considering the group information [41, 42]. The prediction accuracy can be improved due to the inclusion of grouped functional information [43–45]. For example, Tang et al. [45] recently designed a group spike-and-slab Lasso generalized linear model (gsslasso) that combined KEGG pathway information into model fitting and demonstrated that compared with regularization methods (e.g. Lasso), the average gain of prediction accuracy (measured by area under the curve (AUC)) of gsslasso was about 4.5% for sarcoma, 4.6% for ovarian cancer, and about 1.6% for breast cancer by leveraging gene expression data available from the Cancer Genome Atlas (TCGA) [46].

However, how to appropriately include grouped functional information into genetic prediction models is less understood in the literature. Model-averaging methods [47, 48] offer a natural manner to address this problem by averaging the performance of multiple candidate prediction models which can be efficiently constructed based on grouped genetic datasets. Motivated by this, in the present study, we employ a novel model-averaging approach for high dimensional genetic risk prediction while incorporating KEGG pathway information into the model specification. The proposed model-averaging approach selects the optimal weights across candidate models by minimizing a cross validation criterion in a jackknife way. We thus refer to the method as jackknife model averaging prediction (JMAP). We use extensive simulation studies to evaluate the performance of JMAP and compare it with existing methods. Finally, we apply JMAP to four real cancer datasets that are publicly available from TCGA. To construct candidate prediction models, in the present study, we divide genes in terms of the KEGG pathway information [37, 38].

2. Methods and Materials

2.1. Overview of the JMAP Method. We first present an overview of JMAP here; the detailed description of JMAP is shown in Supplementary Materials. Briefly, JMAP consists of two-step model fitting procedures: (i) in the first step, we divide the molecular predictors (e.g., genome-wide gene expressions) into K biological pathways/groups (e.g., KEGG) and build a series of candidate linear prediction models with gene expression measurements available for various groups; we assume that the pathways are predetermined and that the predictors may overlap across different pathways; (ii) in the second step, we look for a suitable weight vector for averaging across the candidate models to perform a pooled prediction. One of the primary features of JMAP is to allow model weights to vary from 0 to 1 but without the limitation that the summation of weights is necessarily equal to one [47, 48]. As we will see, this weight relaxation is important and critical, resulting in a substantial improvement of the prediction accuracy. JMAP has been implemented within an R function freely available at https://github.com/biostatpzeng.

2.2. Simulations and Real Data Applications

2.2.1. Simulation Settings. We next carried out extensive simulations to evaluate the prediction performance of JMAP. To make the simulation settings as real as possible, we used gene expression levels obtained from an existing TCGA dataset of breast cancer (see below for further information about this data). For simplicity, we extracted the expression levels for 6,000 randomly selected genes and 500 breast cancer patients and simulated phenotypes using the following model:

$$\mathbf{y} = \sum_{j=1}^{K} \mathbf{G}_j \boldsymbol{\beta}_j + \mathbf{e}, \mathbf{e} \sim N\left(0, \mathbf{I}_n \sigma_e^2\right), \tag{1}$$

where K is the total number of groups (or pathways); $\mathbf{G}_j$ is an $n \times m_j$ genetic matrix for m_j genes in group j with n the sample size (here n = 500), $\boldsymbol{\beta}_j$ is an m_j-dimensional vector of effects sizes; $\mathbf{I}_n$ is an $n \times n$ identity matrix; and $\mathbf{e}$ is an n-dimensional vector of independently and normally distributed residuals with variance σ_e^2. We considered four scenarios with different group partitions. In scenarios 1–3, genes were sequentially divided into 50, 200, or 300 groups with approximately equal genes per each group; no overlapping of genes existed among groups. In scenario 4, we classified genes into 328 groups in terms of the KEGG pathway information (see below for details); note that, under this case, the number of genes included in each group was not equal and ~21% genes belonged to multiple pathways. Then, following [45], in each scenario, we randomly selected five out of all K groups (K = 50, 200, 300, or 328 as defined above) and generated: (I) the effect sizes $\boldsymbol{\beta}_l$ (l = 1, 2, 3, 4 and 5) in each of the selected groups followed a normal distribution with mean zero and the same variance (say σ_l^2). Under this case, all the genes in the five groups had nonzero effect sizes; (II) unlike case I, here, we assumed that only the genes in the first two groups had nonzero effect sizes and half of the genes in the last three group had nonzero effect sizes; (III) instead of assuming equal proportion of nonzero effect sizes in the last three groups, we set the proportion of nonzero effect sizes to be 80%, 50%, and 20%, respectively; (IV) in this case, we set the proportion of nonzero effect sizes to be 90%, 70%, 50%, 30%, and 10% for the five groups, respectively. The variance parameters σ_l^2 and σ_e^2 were carefully chosen to ensure that $\mathbf{y}$ had unit variance asymptotically, and the phenotypic variance explained (PVE) by genetic component was 0.3, 0.5, or 0.8 in each case, respectively. The effect sizes for the unselected gene groups were set to zero.

2.2.2. Real Data Applications. We now applied JMAP to four cancer datasets publicly available from TCGA [46], including breast cancer (BRCA), colon and rectal cancer (CRC), colon cancer (COAD), and pancreatic cancer (PAAD). We downloaded both the clinical data and RNAseq gene expression levels for those cancers from UCSC Xena (https://xenabrowser.net/). For each cancer, we first merged the clinical data and gene expression levels measured from primary cancer tissue; then, we removed genes with more than 50% zero expressions and standardized the remaining gene expression levels. The used datasets in this study are summarized in Table 1. Following previous studies [12, 38, 49], for the four cancers, we used the age at initial pathologic diagnosis (i.e., onset age) as phenotypes because the age of onset is an important indicator that the cancer is likely more commonly genetic in origin. We quantile-normalized onset age to a standard normal distribution before prediction analysis.

2.3. Model Comparison and Implementation. For the simulated data, the genes were divided into 50, 200, 300, or 328 groups under various scenarios as mentioned before. For the real datasets, we mapped the genes to KEGG pathways by *R* package clusterProfiler (version 3.8.1) after matching gene symbols to Entrez ids [50] and divided the genes into 328 pathway groups. For both simulated and real datasets, following [24], we performed 100 Monte Carlo cross validation (MCCV) data splits by randomly selecting 80% samples as training data and the remaining 20% as test data. We fitted the prediction models in the training data and evaluated the performance in the test data with correlation coefficient (*R*).

As gsslasso was proved to perform better than sparse group Lasso [45]; our competing methods only included Lasso [51], elastic net (ENET) [52], random forest [53], and gsslasso [45]. For both Lasso and ENET, we implemented them via the *R* package glmnet (version 2.0-16), selected the optimal penalty parameters in Lasso and ENET using 100-fold cross validation, and set $\alpha = 0.50$ in ENET as done in [54]. For random forest, we implemented it via the *R* package randomForest (version 4.6-14). For gsslasso, we implemented it via the *R* package BhGLM (version 1.1.0). Following [45], we selected the optimal penalty parameter of gsslasso by setting the slab scale (denoted by s_1) to 1, calculated the accuracy of prediction for a series values for the spike scale (denoted by s_0) (i.e., $s_0 = 0.01 \times m$, $m = 0.1$, 1, 2, . . ., 9), and chose the optimal value for s_0 that resulted in a highest prediction. We solved the quadratic problem in JMAP (Equations (7) and (8) in Supplementary Materials) using the optim function in *R* statistical software. We further contrasted the prediction performance of all other methods with that of JMAP by taking the difference of *R* between the other methods and JMAP. Therefore, an *R* difference below zero suggests worse performance than JMAP.

Table 1: Sample sizes and the number of genes for each cancer in the TCGA dataset used in our analysis.

Phenotypes	Initial gene expression data		Initial clinical data (*N*)	Final data after quality control	
	N	*G*		*N*	*G*
BRCA	1,218	20,531	1,247	1,083	17,675
COAD	329	20,531	551	275	17,493
CRC	434	20,531	736	367	17,510
PAAD	183	20,531	196	178	17,675

Note. N is the sample size and *G* denotes the number of genes. The average number of genes incorporated in each pathway for the seven phenotypes was 65 (ranging from 1 to 1,139), and about 21% genes belonged to multiple pathways. BRCA: breast cancer; CRC: colon and rectal cancer; COAD: colon cancer; PAAD: pancreatic cancer.

3. Results

3.1. Results of the Simulation Studies. The simulation results for the difference of *R* with PVE = 0.3 are shown in Figure 1 with the original *R* values shown in Figure S1. There are 16 combinations presented in Figure 1. Compared with other existing approaches (i.e., Lasso, ENET, random forest, and gsslasso), we find that, except two situations, JMAP performed best or is among the best methods in most of the combinations (14 out of 16). For example, among those 14 settings, JMAP has an average of 0.075 higher prediction accuracy compared with gsslasso, with the difference of *R* ranging from 0.023 to 0.116. In the setting with 200 groups in scenario I (where all the genes in the five groups had nonzero effect sizes), JMAP is better than gsslasso (0.056 higher) and is comparable with random forest, while it behaves slightly worse than Lasso (0.012 lower) and ENET (0.013 lower). In the setting with 300 group in scenario III (where the genes among the first two groups had nonzero effect sizes, but some of the genes in the rest three groups are null with various null proportions), all the four competitive methods (i.e., Lasso, ENET, random forest, and gsslasso) have a higher prediction accuracy relative to JMAP. The simulation results for PVE = 0.5 and 0.8 are displayed in Figures S2–S5 in Supplementary Materials; we observed the similar pattern that JMAP performs better or is as good as other competing methods in most of the simulated settings. We further checked the estimated weights for the candidate models in all the scenarios and found that the weights for the true candidate models (i.e., those with nonzero effect sizes) have much smaller chances to be zero compared with those for the null candidate models and are substantially greater in magnitude (Table S1).

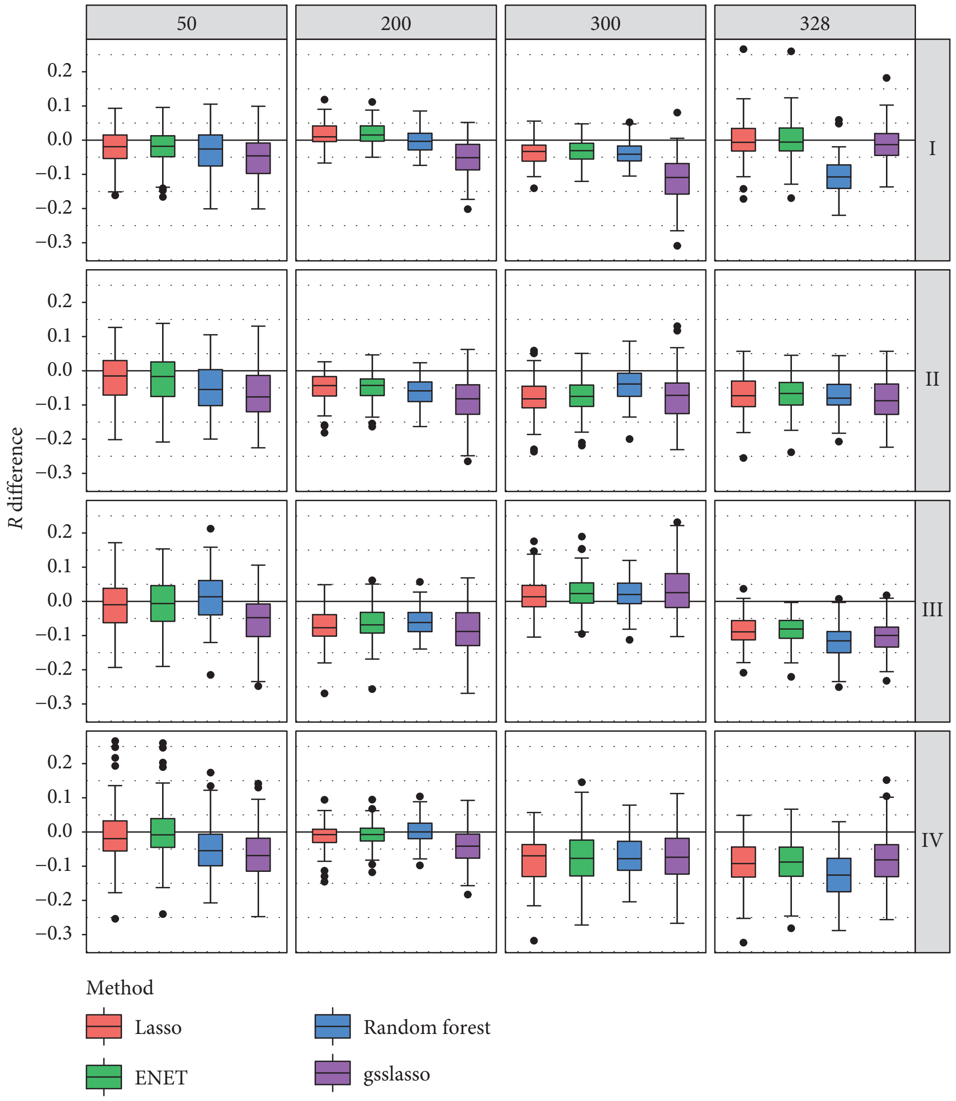

Figure 1: Comparison of predictive performance of four models with JMAP with PVE = 0.3. Performance is measured by *R* difference with respect to JMAP; therefore, a negative value (i.e., values below the horizontal line) indicates worse performance than JMAP. In each setting, five groups with nonzero effect sizes were selected; I represents the settings where all the genes in the five groups had nonzero effect sizes; II represents the settings where only the genes in the first two groups had nonzero effect sizes, and half of the genes in the last three groups had nonzero effect sizes; III represents the settings where the effect sizes of the first two groups were nonzero, and the proportion of nonzero effect sizes in the last three groups was 80%, 50%, or 20%; IV represents the settings where the proportion of nonzero effect sizes in the five groups was 90%, 70%, 50%, 30%, or 10%. The predictive performance was assessed across 100 replicates in each scenario.

3.2. Results of the Real Data Applications. Now, we turn to the real application of the TCGA data (Table 1). The results of *R* differences of other four methods compared with JMAP are presented in Figure 2. Totally, JMAP performs comparably or better compared with the other methods. For example, for the COAD, CRC, and PAAD datasets, JMAP has the highest predictive power, followed by gsslasso. Compared with gsslasso, in these three datasets, the gains of predictive accuracy of JMAP are 0.019, 0.064, and 0.052, respectively. In the PAAD dataset, JMAP is better than Lasso, gsslasso, and ENET, while random forest has the highest prediction accuracy. In the BRCA dataset, except for random forest, the rest of the methods (i.e., Lasso, gsslasso, and ENET) have a higher prediction accuracy compared with JMAP.

4. Discussion

In the present study, we have employed a novel statistical method, JMAP, for genetic prediction and evaluation of complex phenotypes from the publicly available TCGA datasets. Traditionally, the classical model-averaging methods first build a series of candidate models with various degrees of model complexity; then combine all the candidate models together to boost the prediction performance by specifying greater weights onto better models; and require the summation of the model weights is equal to one [47, 55, 56]. However, unlike those previous methods, JMAP relaxes the constraint of summing the weights of candidate models up to one. By removing this restriction and including

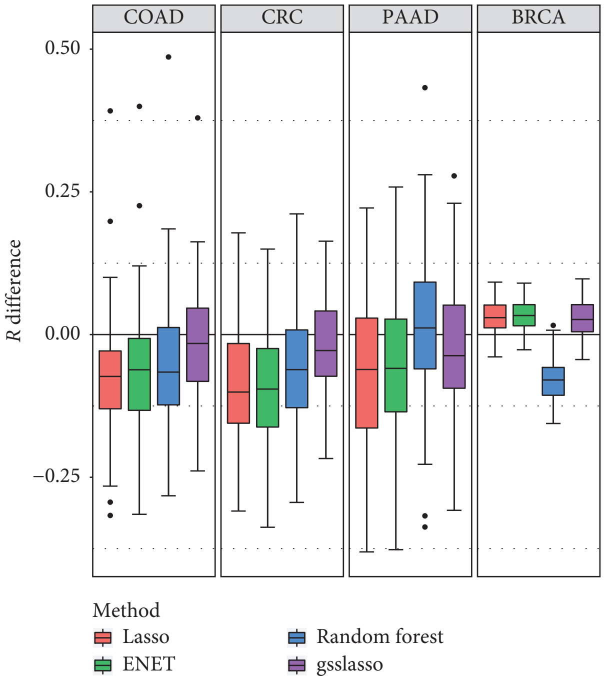

Figure 2: Comparison of predictive performance of four models with JMAP for the four phenotypes from the TCGA datasets. Performance is measured by R difference with respect to JMAP; therefore, a negative value (i.e., values below the horizontal line) indicates worse performance than JMAP. The predictive performance was assessed across 100 MCCV replicates. BRCA: breast cancer; CRC: colon and rectal cancer; COAD: colon cancer; PAAD: pancreatic cancer.

genetic pathway information, as we have demonstrated in the simulations and real data applications, JMAP has shown higher prediction accuracy compared with existing approaches. Furthermore, it is natural to examine whether the weight restriction can be further relaxed to allow them to vary between −1 and 1 [57]. However, we found that this further relaxing may be not beneficial for improving the prediction performance, leading to low accuracy of genetic prediction (Figure S7). Additionally, because each candidate model is fitted with ordinary least squares method which leads to an analytical solution for the effect sizes and because the weight estimation is optimized through a constrained quadratic manner, JMAP is thus computationally efficient and can be easily scalable to the high dimensional genetic risk prediction problem. For example, in the real data applications, it takes only about 3, 3, 110, 15, and 18 seconds on average for Lasso, ENET, random forest, gsslasso, JMAP on the COAD datasets, respectively.

In practice, the candidate models for model averaging are typically established in terms of prior knowledge or expert viewpoints, and the number of the candidate models (i.e., K in our study) is assumed to be uncertain. To address this problem, Ando and Li [48] recently proposed first to partition predictors (equivalent to genes in our study) based on the marginal correlation magnitude between each predictor and the response and then adaptively prepared for candidate model for each partition. This strategy is a flexible way and avoids the requirement of external information, while it may be suboptimal if there is informative prior information that can be utilized. In contrast, in our study, we explicitly preassigned the number of candidate models for JMAP. Indeed, using simulations, we have discovered that JMAP possessed consistently good prediction performance across various candidate model partitions. In our real data applications, we also directly built the candidate models for JMAP based on useful KEGG pathway information which characterizes the biological functions for various sets of genes [37, 38] and can result in each candidate model having unique strength in capturing certain aspects of prediction ability. Applying external informative pathways to establish candidate models in JAMP can lead to at least three benefits: (i) it does not need to search for the appropriate number of candidate models by partitioning all the genes; thus, it is computationally faster; (ii) relying on previously well-validated pathway information, the established candidate models are more biologically meaningful; (iii) finally, the marginal correlation way typically groups a given gene into only one candidate model [48], while in practice, a gene often can be involved in multiple pathways and will be thus included into several candidate models, e.g., in our analysis, about 21% genes can be grouped into at least two pathways. More generally, under the context of model averaging, JAMP can naturally handle the overlapping group structures—a phenomenon that is frequently encountered in pathway-based data analyses [58]. It has been shown that efficiently incorporating the overlapping group structures into model fitting can raise the prediction performance [45]. Hence, JAMP has the potential for further enhancing prediction accuracy. Figures S8 and S9 show the predictive performance of JMAP and MCV2 (i.e., the model-averaging method described in [48], where the candidate models are constructed based on the marginal correlation magnitude between each predictor and the response) for phenotypes from both the simulated and real-life datasets and illustrate the advantage of preassigning the candidate models.

As mentioned before, the greatest feature of JMAP is that the sum of the model weights is equal to one is relaxed. In contrast, the traditional model-averaging approaches often assume that candidate models are equally competitive and thus assign equal weights for all the candidate models. However, in practice, this does not necessarily hold given the fact that only a few pathways are active and the other pathways may have a small or ignorable influence on complex phenotypes. Furthermore, as shown in the simulations and real data applications, relaxing the weights limitation in JMAP allows to put more weights on candidate models that were constructed for possibly active pathways, potentially increasing the prediction performance. Theoretically, the benefit of relaxing the weights limitation in model-averaging approaches has been proved in [48].

It is worth noting that in the candidate model of JMAP, the least squares estimate in Equation (2) (Supplementary Materials) is ill-conditional when the number of genetic markers is larger than the sample size for some genes. For example, in our analysis, there are 5.5% and 5.2% pathways with the number of genes greater than the sample sizes for the

PAAD and COAD datasets, respectively. Under this situation, regularization methods (e.g., Lasso) can be applied to each candidate model [59]; however, doing this can lead to substantial increase in computational time because the simple closed-form solution cannot be available for candidate model. In the present study, by borrowing the idea of ridge regression [60, 61], we have attempted to add a nonnegative constant δ into the estimates, i.e., replacing $\mathbf{G}_j^T\mathbf{G}_j$ with $\mathbf{G}_j^T\mathbf{G}_j + \delta\mathbf{I}_{m_j}$ (Equation (2) in the Supplementary Materials). In our paper, we primarily set δ to be one and found that JMAP is robust with regard to various values of δ with simulations (Figure S10). We emphasize that this is an ad hoc modification which has no clear theoretical foundation. Further investigation of JMAP under the context that the dimension of candidate model is larger than the sample size is an important and interesting topic and is our next research direction.

Finally, the current version of JMAP described in our study is constructed only for continuous phenotypes. Extending model averaging from linear to nonlinear regression under the high dimensional situations was recently investigated [57]. However, although not mentioned, an explicit model assumption in their study is that the number of the predictors in each candidate generalized linear model should be much less than the sample size to ensure the estimates can be identifiable. Therefore, their methods cannot be applied to our case where the number of the genes for some candidate models is easy to be greater than the sample size as mentioned before. Thus, in our real data application, we had to directly fit linear candidate models for binary phenotypes by treating them as continuous values following previous studies [21–23, 25]. Theoretically, modeling binary data with linear models can be justified by the fact that the linear model can be viewed as a first order Taylor approximation to the generalized linear model, and this approximation is accurate when the effect size is weak and small [21]—a condition which generally satisfies because it has been shown that most complex phenotypes are polygenic and are influenced by many genetic variants with small effect sizes [7]. Nevertheless, extending the JMAP model for application to noncontinuous phenotypes in high dimensional prediction problems warrants more explorations.

Data Availability

The TCGA data are publicly available from https://xenabrowser.net/. The BhGLM software is available from http://github.com/nyiuab/BhGLM. The glmnet package is available from https://cran.r-project.org/web/packages/glmnet/index.html. Random forest software is available from https://cran.r-project.org/web/packages/randomForest/index.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

PZ and SH conceived and designed the experiment; XY and PZ cleared up, analyzed, and interpreted the datasets; and XY, LX, and PZ wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors acknowledge the contributions of TCGA Research Network for the public availability of the cancer datasets used in our paper. This study was supported by the National Natural Science Foundation of Jiangsu Province (BK20181472), the Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), the China Postdoctoral Science Foundation (2018M630607), The Jiangsu QingLan Research Project for Outstanding Young Teachers, The Postdoctoral Science Foundation of Xuzhou Medical University, the National Natural Science Foundation of China (81402765), the Statistical Science Research Project from National Bureau of Statistics of China (2014LY112), Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJKY19_2129), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) for Xuzhou Medical University.

Supplementary Materials

A detailed description for the proposed JMAP approach. Briefly, JMAP is a novel model-averaging based genetic risk prediction approach that can incorporate the group biological information of genetic alterations into prediction modeling. It consists of two-step model fitting procedures: (1) construct candidate models and (2) optimize the model weights. (*Supplementary Materials*)

References

[1] G. R. Abecasis, G. R. Abecasis, D. Altshuler et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[2] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 415–425, 2010.

[3] M. L. Metzker, "Sequencing technologies-the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[4] P. AC't Hoen, M. R. Friedländer, J. Almlöf et al., "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories," *Nature Biotechnology*, vol. 31, no. 11, pp. 1015–1022, 2013.

[5] D. Altshuler, M. J. Daly, and E. S. Lander, "Genetic mapping in human disease," *science*, vol. 322, no. 5903, pp. 881–888, 2008.

[6] J. MacArthur, E. Bowler, M. Cerezo et al., "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic acids research*, vol. 45, no. D1, pp. D896–D901, 2016.

[7] P. M. Visscher, N. R. Wray, Q. Zhang et al., "10 years of GWAS discovery: biology, function, and translation," *American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, 2017.

[8] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common

diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[9] C. Fuchsberger, J. Flannick, T. M. Teslovich et al., "The genetic architecture of type 2 diabetes," *Nature*, vol. 536, no. 7614, pp. 41–47, 2016.

[10] C. J. Willer, E. M. Schmidt, S. Sengupta et al., "Discovery and refinement of loci associated with lipid levels," *Nature Genetics*, vol. 45, no. 11, pp. 1274–1283, 2013.

[11] W. van Rheenen, A. Shatunov, A. M. Dekker et al., "Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis," *Nature Genetics*, vol. 48, no. 9, pp. 1043–1048, 2016.

[12] A. Gusev, H. Won, N. Mancuso et al., "Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights," *Nature Genetics*, vol. 50, no. 4, pp. 538–548, 2018.

[13] A. Gusev, A. Ko, H. Shi et al., "Integrative approaches for large-scale transcriptome-wide association studies," *Nature Genetics*, vol. 48, no. 3, pp. 245–252, 2017.

[14] L. Wu, W. Shi, J. Long et al., "A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer," *Nature Genetics*, vol. 50, no. 7, pp. 968–978, 2018.

[15] R. Makowsky, N. M. Pajewski, Y. C. Klimentidis et al., "Beyond missing heritability: prediction of complex traits," *PLoS Genetics*, vol. 7, no. 4, article e1002051, 2011.

[16] G. de los Campos, D. Gianola, and D. B. Allison, "Predicting genetic predisposition in humans: the promise of whole-genome markers," *Nature Reviews Genetics*, vol. 11, no. 12, pp. 880–886, 2010.

[17] N. Chatterjee, J. Shi, and M. García-Closas, "Developing and evaluating polygenic risk prediction models for stratified disease prevention," *Nature Reviews Genetics*, vol. 17, no. 7, pp. 392–406, 2016.

[18] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J.-H. Park, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics*, vol. 45, no. 4, pp. 400–405, 2013.

[19] J. S. Zager, B. R. Gastman, S. Leachman et al., "Performance of a prognostic 31-gene expression profile in an independent cohort of 523 cutaneous melanoma patients," *BMC Cancer*, vol. 18, no. 1, p. 130, 2018.

[20] Y. Jiang, W. Mei, Y. Gu et al., "Construction of a set of novel and robust gene expression signatures predicting prostate cancer recurrence," *Molecular Oncology*, vol. 12, no. 9, pp. 1559–1578, 2018.

[21] X. Zhou, P. Carbonetto, and M. Stephens, "Polygenic modeling with Bayesian sparse linear mixed models," *PLoS Genetics*, vol. 9, no. 2, article e1003264, 2013.

[22] G. Moser, S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher, "Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model," *PLoS Genetics*, vol. 11, no. 4, article e1004969, 2015.

[23] O. Weissbrod, D. Geiger, and S. Rosset, "Multikernel linear mixed models for complex phenotype prediction," *Genome Research*, vol. 26, no. 7, pp. 969–979, 2016.

[24] P. Zeng and X. Zhou, "Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models," *Nature Communications*, vol. 8, no. 1, p. 456, 2017.

[25] D. Speed and D. J. Balding, "MultiBLUP: improved SNP-based prediction for complex traits," *Genome Research*, vol. 24, no. 9, pp. 1550–1557, 2014.

[26] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, and T. Aittokallio, "Regularized machine learning in the genetic prediction of complex traits," *PLoS Genetics*, vol. 10, no. 11, article e1004754, 2014.

[27] E. R. Gamazon, K. P. Shah, H. E. Wheeler et al., "A gene-based association method for mapping traits using reference transcriptome data," *Nature Genetics*, vol. 47, no. 9, pp. 1091–1098, 2015.

[28] T. H. Pers, J. M. Karjalainen, Y. Chan et al., "Biological interpretation of genome-wide association studies using predicted gene functions," *Nature Communications*, vol. 6, no. 1, p. 5890, 2015.

[29] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, 2011.

[30] P. Zeng, Y. Zhao, J. Liu et al., "Likelihood ratio tests in rare variant detection for continuous phenotypes," *Annals of Human Genetics*, vol. 78, no. 5, pp. 320–332, 2014.

[31] P. Zeng, T. Wang, and S. Huang, "Cis-SNPs set testing and PrediXcan analysis for gene expression data using linear mixed models," *Scientific Reports*, vol. 7, no. 1, article 15237, 2017.

[32] H. K. Finucane, B. Bulik-Sullivan, A. Gusev et al., "Partitioning heritability by functional annotation using genome-wide association summary statistics," *Nature Genetics*, vol. 47, no. 11, pp. 1228–1235, 2015.

[33] A. Gusev, S. H. Lee, G. Trynka et al., "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases," *American Journal of Human Genetics*, vol. 95, no. 5, pp. 535–552, 2014.

[34] W. Pan, I.-Y. Kwak, and P. Wei, "A powerful pathway-based adaptive test for genetic association with common or rare variants," *American Journal of Human Genetics*, vol. 97, no. 1, pp. 86–98, 2015.

[35] K. Wang, M. Li, and M. Bucan, "Pathway-based approaches for analysis of genomewide association studies," *American Journal of Human Genetics*, vol. 81, no. 6, pp. 1278–1283, 2007.

[36] H. Zhong, X. Yang, L. M. Kaplan, C. Molony, and E. E. Schadt, "Integrating pathway analysis and genetics of gene expression for genome-wide association studies," *American Journal of Human Genetics*, vol. 86, no. 4, pp. 581–591, 2010.

[37] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research*, vol. 42, no. D1, pp. D199–D205, 2014.

[38] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2015.

[39] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, p. 140, 2007.

[40] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, article e1000217, 2008.

[41] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[42] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," 2010, https://arxiv.org/abs/1001.0736.

[43] Y. Hu, Q. Lu, R. Powles et al., "Leveraging functional annotations in genetic risk prediction for human complex diseases," *PLoS Computational Biology*, vol. 13, no. 6, article e1005589, 2017.

[44] Y. Hu, Q. Lu, W. Liu, Y. Zhang, M. Li, and H. Zhao, "Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction," *PLoS Genetics*, vol. 13, no. 6, article e1006836, 2017.

[45] Z. Tang, Y. Shen, Y. Li et al., "Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information," *Bioinformatics*, vol. 34, no. 6, pp. 901–910, 2018.

[46] K. A. Hoadley, C. Yau, T. Hinoue et al., "Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, pp. 291.e6–304.e6, 2018.

[47] B. E. Hansen and J. S. Racine, "Jackknife model averaging," *Journal of Econometrics*, vol. 167, no. 1, pp. 38–46, 2012.

[48] T. Ando and K.-C. Li, "A model-averaging approach for high-dimensional regression," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 254–265, 2014.

[49] K. L. Huang, R. J. Mashl, Y. Wu et al., "Pathogenic germline variants in 10,389 adult cancers," *Cell*, vol. 173, no. 2, pp. 355.e14–370.e14, 2018.

[50] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *Omics: A Journal of Integrative Biology*, vol. 16, no. 5, pp. 284–287, 2012.

[51] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[52] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[53] R. Diaz-Uriarte and S. A. de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[54] P. Zeng, X. Zhou, and S. Huang, "Prediction of gene expression with cis-SNPs using mixed models and regularization methods," *BMC Genomics*, vol. 18, no. 1, p. 368, 2017.

[55] A. T. K. Wan, X. Zhang, and G. Zou, "Least squares model averaging by Mallows criterion," *Journal of Econometrics*, vol. 156, no. 2, pp. 277–283, 2010.

[56] X. Zhang, G. Zou, and H. Liang, "Model averaging and weight choice in linear mixed-effects models," *Biometrika*, vol. 101, no. 1, pp. 205–218, 2014.

[57] T. Ando and K.-c. Li, "A weight-relaxed model averaging approach for high-dimensional generalized linear models," *Annals of Statistics*, vol. 45, no. 6, pp. 2654–2679, 2017.

[58] M. Silver, G. Montana, and Alzheimer's Disease Neuroimaging Initiative, "Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, pp. 1–43, 2012.

[59] B. Lin, Q. Wang, J. Zhang, and Z. Pang, "Stable prediction in high-dimensional linear models," *Statistics and Computing*, vol. 27, no. 5, pp. 1401–1412, 2017.

[60] A. E. Hoerl and R. W. Kennard, "Ridge regression: applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.

[61] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2nd edition, 2009.