BioTech

MDPI

*Article*

# Integrating Multi-Omics Data for Gene-Environment Interactions

**Yinhao Du, Kun Fan, Xi Lu and Cen Wu ***

Department of Statistics, Kansas State University, Manhattan, KS 66506, USA; ydu@ksu.edu (Y.D.); kfan@ksu.edu (K.F.); xilu@ksu.edu (X.L.)
* Correspondence: wucen@ksu.edu

**Abstract:** Gene-environment (G×E) interaction is critical for understanding the genetic basis of complex disease beyond genetic and environment main effects. In addition to existing tools for interaction studies, penalized variable selection emerges as a promising alternative for dissecting G×E interactions. Despite the success, variable selection is limited in terms of accounting for multidimensional measurements. Published variable selection methods cannot accommodate structured sparsity in the framework of integrating multiomics data for disease outcomes. In this paper, we have developed a novel variable selection method in order to integrate multi-omics measurements in G×E interaction studies. Extensive studies have already revealed that analyzing omics data across multi-platforms is not only sensible biologically, but also resulting in improved identification and prediction performance. Our integrative model can efficiently pinpoint important regulators of gene expressions through sparse dimensionality reduction, and link the disease outcomes to multiple effects in the integrative G×E studies through accommodating a sparse bi-level structure. The simulation studies show the integrative model leads to better identification of G×E interactions and regulators than alternative methods. In two G×E lung cancer studies with high dimensional multi-omics data, the integrative model leads to an improved prediction and findings with important biological implications.

**Keywords:** Gene-environment (G×E) interactions; integrated analysis; multidimensional data; high-dimensional variable selection



**Citation:** Du, Y.; Fan, K.; Lu, X.; Wu, C. Integrating Multi-Omics Data for Gene-Environment Interactions. *BioTech* **2021**, *10*, 3. https://doi.org/10.3390/biotech10010003

Received: 24 December 2020
Accepted: 22 January 2021
Published: 29 January 2021



## 1. Introduction

Gene-environment interactions reveal how the changes in environmental exposures mediate the contribution of genetic factors in order to influence the variations in disease traits, which makes it critical in understanding the comprehensive genetic architecture of complex diseases [1,2]. Traditionally, G×E interaction studies have mainly been conducted within the framework of genetic association studies in order to hunt down the important main and interaction effects that are associated with the disease phenotypes [3,4].

Most of the existing G×E studies are one-dimensional, in that the interactions between environmental factors and one type of genetic factor (such as gene expression or SNPs) have been considered. In the multi-omics era, there is a pressing need to account for multi-platform measurements in G×E studies. Consider a G×E analysis with environmental factors and gene expression (GE) as the G factors. In addition, DNA methylation (DM) and copy number alterations (CNA), which are the regulators of the genetic factors, are also available. A typical G×E analysis only focuses on the interaction effects that involve the G factor (GE) and ignores its regulators, losing the extra power of elucidating the genetic basis of complex disease while using multi-level omics data.

Integrating multi-omics data for prognostic outcomes has mainly been conducted using parallel and horizontal integration strategies [5]. With the parallel integration, different types of omics measurements are treated equally, and important associations between these measurements and the prognostic outcome are identified in a joint model. On the other hand, the hierarchical integration fully accounts for the regulatory information by

accommodating the indirect effects of regulators, such as DM and CNA, on the prognostic outcomes that are mediated through GEs. Meanwhile, the direct effects of regulators on the cancer outcomes, which have not been captured by GEs through other mechanisms, such as post-transcriptional regulations, should also been taken into consideration.

Given the availability of multi-omics features, the major limitation of existing G×E interaction studies lies in the incapability of integrating regulators in the interaction model under prognostic outcomes, which has motivated us to develop a two stage integrative model for G×E interaction analysis while using multi-level cancer omics data. At the first stage, the sparse regulatory relationship has been determined through penalization, where the linear regulatory modelling [6], or LRM, has been adopted in order to identify the sets of regulators that influence the sets of GEs, as well as the residuals of gene expression and residuals of regulators that cannot be captured by the LRMs. At the second stage, the LRMs and both types of residuals are treated as direct effects on cancer outcomes in the G×E model, and penalization has been conducted in order to identify the important main and interaction effects.

In the past decade, the effectiveness of regularization for G×E interaction studies has been increasingly witnessed [7]. Extension of the technique for an integrated interaction study is not trivial. Our method significantly advances from existing integration studies not tailored for interaction structures and interaction analysis ignoring the multidimensional omics measurements. Extensive simulation studies, have been performed to demonstrate the advantage of the proposed method over multiple alternatives. Our method leads to main and interaction effects with sensible biological implications and improved prediction performance in two case studies of the lung cancer data (LUSC and LUAD) from TCGA.

## 2. Method

Let $Y_{n\times 1}$ denote cancer outcome, $E_{n\times q} = (E_1, \cdots, E_q)$ denote the $q$ environmental factors, $G_{n\times p_g} = (G_1, \cdots, G_{p_g})$ denote the $p_g$ gene expressions, and $R_{n\times p_r} = (R_1, \cdots, R_{p_r})$ denote the $p_r$ regulators. Suppose that we have two measurements for the regulators, $p_{r_1}$ DM and $p_{r_2}$ CNA, then we can obtain $R_{n\times p_r}$ by stacking the measurements together with $p_r = p_{r_1} + p_{r_2}$. Next, we describe the overall analysis framework and integrative model.

### 2.1. Analysis Framework

First, consider a G×E model in the multi-omics scenario, where the regulators of the $G$ factors are also included, in addition to the main and interaction effects.

$$Y = \sum_{k=1}^{q} \alpha_k E_k + \sum_{j=1}^{p_g} \left( \beta_j G_j + \sum_{k=1}^{q} \eta_{jk} G_j E_k \right) + \sum_{t=1}^{p_r} \gamma_t R_t + \epsilon, \quad (1)$$

where $\alpha_k$, $\beta_j$, and $\eta_{jk}$ are the regression coefficients for the $k$th environmental factor, $j$th gene expression and their interactions, respectively. Besides, $\gamma_t$ is the regression coefficient for the $t$th regulator and $\epsilon$ is the random error.

Model (1) shares the spirit of parallel integration by treating the genetic factor and its regulators equally. Although such a strategy has shown to be effective in several studies, a more attractive alternative is to conduct vertical integration via accounting for the regulatory information among the different levels of omics measurements [5]. Typically, integrating multi-omics data in a main effect model with prognostic outcomes consists of two steps. At the first step, the sparse regulatory relationship can be identified, which leads to gene expressions that are modulated and not modulated by regulators, which can then be linked to clinical outcomes at the second step [6,8]. Specifically, Zhu et al. [6] proposed the linear regulatory model (LRM) to pinpoint the set of regulators that affect the corresponding set of GEs. Subsequently, the clinical model incorporates the GEs, residual GEs, and residual regulators. In this study, we extend the LRM to investigate the G×E

interactions in the presence of multi-level omics measurements. In particular, the prognostic model at the second stage consists of : (1) a low dimensional environmental factors; (2) regulated GEs in the form of LRMs from the first stage and their interactions with those environmental factors; (3) Residual GEs and their interactions with environmental factors; and, (4) the residual effects of regulators.

### 2.2. Stage 1: The Linear Regulatory Model (LRM)

Denote $g = (g_1, \cdots, g_{p_g})$ as the $p_g$ gene expressions and denote $r = (r_1, \cdots, r_{p_r})$ as the $p_r$ regulators. The LRM can be expressed as

$$E(gV_{p_g \times L}|r) = a_{1\times L} + rU_{p_r \times L}, \tag{2}$$

where $a$ is the intercept, $V = (v_1, \cdots, v_L)$ and $U = (u_1, \cdots, u_L)$ both contain $L$ columns of loading vectors ($v_l$ and $u_l$ for $l \in \{1, \cdots, L\}$). Denote $L$ as the total number of LRMs. Here, we assume $U$ and $V$ have orthogonal columns, such that $u_l \perp u_{l^\top}$, $v_l \perp v_{l^\top}$, for $l \neq l^\top$. With this assumption, no overlap between gene expressions and regulators exists in LRM. We expect that different LRMs represent different regulated relationship between gene expressions and regulators [9]. In addition, $v_l$ and $u_l$ are assumed as sparse loading vectors, as only a small number of gene expressions is regulated by, at most, a small number of regulators [10].

For the $j$th gene expression, $j = 1, \cdots, p_g$, we right multiply $V^\top$ to both sides in order to simplify Equation (2). Afterwards, the LRM can be formulated as a regression model with response variable $g_j$ and predictors $r$:

$$E(g_j) = a_j^\top + r\theta_j, \text{ for } j = 1, \cdots, p_g, \tag{3}$$

where $a_j^\top$ is an intercept and $\theta_j$ is the regression coefficient vector. Equation (3) indicates that one gene expression is regulated by a number of regulators. We impose sparsity on $\theta_j$ through penalization to identify a sparse regulatory relationship. Subsequently, the penalized regression model can be written as

$$\frac{1}{2n}\left\|g_j - a_j^\top - r\theta_j\right\|_2^2 + \lambda|\theta_j|, \text{ for } j = 1, \cdots, p_g, \tag{4}$$

where $\lambda$ is the tuning parameter. The LASSO is adopted for its computational simplicity and satisfactory performance [11]. Equation (4) leads to a regularized estimate of $\theta_j$, which indicates that one gene expression is regulated by a limited amount of regulators.

Next, we further investigate the relationship between sets of gene expressions and regulators through singular value decomposition (SVD). The regression model (3) can be collectively written as

$$E(g) = \mathbf{a}^\top + r\Theta_{p_r \times p_g} \tag{5}$$

where $\mathbf{a}^\top$ is the vector of the intercept, $g_{1\times p_g} = (g_1, \cdots, g_{p_g})$, $r_{1\times p_r} = (r_1, \cdots, r_{p_r})$, and $\Theta_{p_r \times p_g} = (\theta_1, \cdots, \theta_{p_g})$ is the transition matrix. The SVD is performed on the transition matrix in order to separate the regression coefficients representing gene expression and regulators:

$$\Theta = UDV^\top = (u_1, \cdots, u_L)D(v_1, \cdots, v_L)^\top \tag{6}$$

where $D = \text{diag}(d_1, \cdots, d_L)$ is a diagonal matrix with $L$ diagonal elements. The diagonal matrix $D$ can account for the dissimilarity among loading vectors in terms of different scaling factors. Subsequently, we can obtain the estimated coefficients for gene expression and regulators by decomposing the estimated transition matrix $\hat{\Theta}$. Under the sparse condition, one gene expression is only regulated by a few of regulators, and one regulator affects a few of gene expressions [10]. In order to impose sparsity, we adopt the sparse

SVD method that was developed by Lee et al. (2010) [12], where sparse singular vectors that correspond to the largest singular values are recursively obtained. Consider the first largest singular value $(d_1, u_1, v_1)$, then the regularized sparse SVD can be expressed as

$$\frac{1}{2n}\left\|\hat{\Theta} - d_1 u_1 v_1\right\|_F^2 + \lambda|d_1 u_1| + \lambda|d_1 v_1| \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm. Tuning parameter $\lambda$ is the same for $u_1$ and $v_1$ for computation efficiency. Here $d_1$ is treated as the scaling factor. After estimating $(d_1, u_1, v_1)$, we update $\hat{\Theta} = \hat{\Theta} - \hat{d}_1 \hat{u}_1 \hat{v}_1^\top$ and recursively update $(d_l, u_l, v_l)$, for $l = 2, \cdots, L$ in a similar manner. With sparse SVD, we can decompose the coefficient and impose sparsity on $p_z$ and $p_x$ for every LRM. The standard LASSO is not applicable within the current LRM formulation, since the shrinkage has been imposed on scaled singular vectors.

### *2.3. Stage 2: The Penalized G×E Interaction Model*

Now, we integrate multiomics measurements for G×E interactions. The regulated GEs, residual GEs, as well as residual regulators can be obtained through LRMs. The $G$ factors are represented by regulated GEs and residual GEs, which are involved in the interaction with dimensional environmental factors. The partition of gene expressions into regulated and non-regulated components proceeds, as follows. The $L$ sets of regulated gene expressions ($GV$) are equivalent to the corresponding sets of regulators ($RU$). We include the $L$ sets of regulated GEs ($GV$) in the G×E model , since gene expressions are more directly related to cancer outcomes. The residual GEs, i.e., the non-regulated GEs that cannot be captured by LRMs, is denoted as $\tilde{G}_{n\times p_g}$. The G factors, consisting of both $GV$ and $\tilde{G}$, interact with $q$ environmental factors. Denote $W_j = (G_j V_j, G_j V_j E_1, \cdots, G_j V_j E_q, \tilde{G}_j, \tilde{G}_j E_1, \cdots, \tilde{G}_j E_q)$, $(j = 1, \cdots, p_g)$. Subsequently, $W_j$ corresponds to the interaction with respect to the $j$th GE. We only consider the main effect of residual regulators, because the influences of regulators on cancer outcomes are mostly mediated by gene expressions, and investigating its interactions with environmental factors is not of interest.

The quantifications of the residuals $\tilde{G}$ and $\tilde{R}$ are conducted through perpendicular projection operation. Because both can be calculated in the same manner, we take $\tilde{G}$ as an example. For the $j$th gene expression, define $S_j$ as the set of all LRMs that contains the $j$ th gene expression. If $S_j$ is empty, then the $j$th gene expression is not regulated, which results in $\tilde{G}_j = G_j$. If $S_j$ is not empty, we denote $V_{S_j}$ as the sub-matrix of $V$ that only contains columns (LRMs) of the $j$th gene expression. Following the perpendicular projection operation, we calculate the residual as $\tilde{G}_j = (I - GV_{S_j}((GV_{S_j})^\top (GV_{S_j}))^{-1}(GV_{S_j})^\top)G_j$, which is the projection of $G_j$ onto the orthogonal space of $GV_{S_j}$.

Consider $n$ subjects, $p_g$ gene expressions, and $L$ LRMs. Subsequently, all of the main and interaction effects can be collectively written as

$$W = (GV, GVE_1, \cdots, GVE_q, \tilde{G}, \tilde{G}E_1, \cdots, \tilde{G}E_q) = (X_1, X_2),$$

where $X_1 = (GV, GVE_1, \cdots, GVE_q)$ denotes the main effects of regulated GEs and their interactions with the environmental factors. Similarly, the effects that correspond to residual GEs are defined as $X_2 = (\tilde{G}, \tilde{G}E_1, \cdots, \tilde{G}E_q)$. Subsequently, we consider the following penalized regression models for G×E interactions:

$$\frac{1}{2n}\left\|Y - \sum_{k=1}^{q}\alpha_k E_k - \sum_{l=1}^{L} X_{1l} b_{1l} - \sum_{j=1}^{p_g} X_{2j} b_{2j} - \sum_{t=1}^{p_r}\gamma_t \tilde{R}_t\right\|_2^2 + \sum_{l=1}^{L} P_1\left(b_{1l}; \lambda_1\right) + \sum_{j=1}^{p_g} P_2\left(b_{2j}; \lambda_2\right) + \sum_{t=1}^{p_r} P_3\left(\gamma_t; \lambda_3\right) \tag{8}$$

where $X_{1l} = (GV_l, GV_l E_1, \cdots, GV_l E_q)$, $(l = 1, \cdots, L)$ represents the $l$th LRM and its interaction with $q$ environmental factors, and $X_{2j} = (\tilde{G}_j, \tilde{G}_j E_1, \cdots, \tilde{G}_j E_q)$, $(j = 1, \cdots, p_g)$

denotes the main and interaction effects with respect to the $j$th residual GEs. Here, $b_{1l}$ and $b_{2j}$ are the corresponding regression coefficients for $X_{1l}$ and $X_{2j}$. $\gamma_t$ is the coefficients for $\tilde{R}_t$ $(t = 1, \cdots, p_r)$, the residual of regulators. $P_i(\cdot; \lambda_i)$, $(i = 1, 2, 3)$, is the penalty function with $\lambda_i$ as the tuning parameter to impose sparsity. The three tuning parameters are set as the same because regression coefficients from the three components are on a similar scale, and different tunings dramatically increase the computational cost. Regularized identification in G×E interaction studies demands tailored penalty functions [7]. For instance, $b_{1l}$ stands for all the main and interaction effects with respect to the $l$th LRM. The selection of $b_{1l}$ on the group levels determines whether the $l$th LRM has any effect at all. If so, then selection of the individual effects within the group further determines the main and/or interactions that are associated with the cancer outcome. Therefore, penalized selection should accommodate the bi-level (or sparse group) structure. To be consistent with the analysis in stage 1, we still adopt LASSO as the baseline penalty function. Specifically, we have

$$P_1(b_{1l}; \lambda_1) = \lambda_1 \|b_{1l}\|_2 + \lambda_1 \sum_{k=1}^{q+1} |b_{1lk}|,\ P_2(b_{2j}; \lambda_2) = \lambda_2 \|b_{2j}\|_2 + \lambda_2 \sum_{k=1}^{q+1} |b_{2jk}|,$$

where $P_1(b_{1l}; \lambda_1)$ and $P_2(b_{2j}; \lambda_2)$ are sparse group LASSO. The L1 norm and L2 norm ($\|\cdot\|_2$) result in penalized identification on the individual and group level, respectively. The sparse group regularization has been adopted for the bi-level selection of main and interaction effects on the individual and group level simultaneously. Its advantage over LASSO in G×E studies has been demonstrated in multiple studies [7]. A corresponding price paid is computational cost, as different bi-level regularization usually demands different tunings. Because we only consider the main effect of residuals of regulators, the L1 norm penalty is adopted for $\gamma_t$ $(t = 1, \cdots, p_r)$. Because the number of environmental factors is usually low, the selection of them is not of interest. They are pre-determined with evidence of being associated with cancer from previous studies. The proposed regularization respects a weak hierarchy between main and interaction effects as the penalty has not been imposed on the environmental main effects. Accordingly, once an interaction effect is selected, at least one of the two corresponding main effects will be in the model.

### *2.4. Computation*

The Equation (8) can be expressed as:

$$\frac{1}{2n}\|Y - E\alpha - X_1 b_1 - X_2 b_2 - \tilde{R}\gamma\|_2^2 + P_1(b_1; \lambda_1) + P_2(b_2; \lambda_2) + P_3(\gamma; \lambda_3) \tag{9}$$

where $\alpha_{q\times 1} = (\alpha_1, \cdots, \alpha_q)^\top$ is the coefficient vector for $q$ environmental factors, $b_{1_{L(q+1)\times 1}} = (b_{1_1}, \cdots, b_{1_L})^\top$ and $b_{2_{p_g(q+1)\times 1}} = (b_{2_1}, \cdots, b_{2_{p_g}})^\top$ are the coefficient vectors for the main and interaction effects of the regulated and residual GEs, respectively. In addition, $\gamma_{p_r\times 1} = (\gamma_1, \cdots, \gamma_{p_r})\top$ is the coefficient vector for residual regulators.

The integrative analysis consists of two steps. In the first step, the loading matrices U and V are estimated through the construction of LRMs. The $j$th column of $\hat{\Theta}$, which is denoted as $\hat{\theta}_j$, $(j = 1, \cdots, p_g)$, is estimated by minimizing Equation (4). For $l = 1, \cdots, L$, the singular vectors that correspond to the largest singular values, $(\hat{u}_l, \hat{v}_l, \hat{d}_l)$, are conducted through the rank-1 sparse SVD on $\hat{\Theta}$. The rank-1 sparse SVD is recursively performed for $l = 1, \cdots, L$, by updating $\hat{\Theta}^{(l+1)} = \hat{\Theta}^{(l)} - \hat{u}_l \hat{d}_l \hat{v}_l^\top$ at each $l$. In the second step, the shrinkage estimate of the regression coefficients can be obtained in the G×E model, where $GV$, $RU$, residuals of gene expressions ($\tilde{G}$), and residuals of regulators ($\tilde{R}$) are calculated accordingly. At the $k$th iteration, the vector of estimated regression coefficients for all of the environmental factors is computed by $\hat{\alpha}^{(k+1)} = (E^{(k)^\top} E^{(k)})^{-1} E^{(k)^\top} (Y - X_1 \hat{b}_1^{(k)} - X_2 \hat{b}_2^{(k)} - \tilde{R}\hat{\gamma}^{(k)})$. Given $\hat{\alpha}^{(k+1)}$ fixed at the current estimate, we obtain $(\hat{b}_1^{(k+1)}, \hat{b}_2^{(k+1)}, \hat{\gamma}^{(k+1)})$

by minimizing Equation (9). The iteration stops until convergence. Algorithm 1 shows the outline of algorithm:

**Algorithm 1** The Integrative analysis for G×E Interaction

**Step 1:** Estimate the loading matrices of LRMs U and V: construct LRMs.

(a) For $j = 1, \cdots, p_g$, obtain $\hat{\theta}_j$ by minimizing Equation (4). Then the estimate $\hat{\Theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_{p_g})$.

Initialize $l = 1$.

**for** $l = 1, \cdots, L$ **do**

(b) Apply rank-1 sparse SVD on $\hat{\Theta}$ to obtain the singular vectors corresponding to largest singular values $(u_l, v_l, d_l)$.

(c) Update $\hat{\Theta}^{(l+1)} = \hat{\Theta}^{(l)} - u_l d_l v_l^\top$.

(d) $l = l + 1$.

**end for**

**Step 2:** Estimate regression coefficients $\alpha, b_1, b_2, \gamma$: construct the penalized G×E interaction model.

(a) Calculate $GV$, $RU$, $\tilde{G}$ and $\tilde{R}$.

Initialize $\hat{b}_1^{(0)} = \hat{b}_2^{(0)} = \hat{\gamma}^{(0)} = 0$.

At the $(k+1)$th iteration.

**repeat**

(b) Compute $\hat{\alpha}^{(k+1)} = (E^{(k)\top} E^{(k)})^{-1} E^{(k)\top} (Y - X_1 \hat{b}_1^{(k)} - X_2 \hat{b}_2^{(k)} - \tilde{R}\hat{\gamma}^{(k)})$.

(c) Obtain $(\hat{b}_1^{(k+1)}, \hat{b}_2^{(k+1)}, \hat{\gamma}^{(k+1)})$ by minimizing Equation (9) through bi-level selection.

**until** convergence

LASSO is adopted in order to conduct the selection of important LRMs from the first stage. At the second stage, a sparse group LASSO has been formulated to accommodate the identification of main and interaction effects on both the group and individual level. We conjecture that other penalization methods, such as adaptive LASSO [13], SCAD [14], and MCP [15], are also applicable in our framework. For example, MCP can be adopted in order to identify sparse regulatory relationship from the first stage, and a sparse group MCP is also tailored for the identification of important G×E interactions in the clinical model. We do not compare the performances of different baseline penalization methods within our framework, as it is not the main interest here.

At the first step, we only use one tuning parameter $\lambda$ for conducting sparse SVD, due to the similarity in scales between GE and its regulators. The three tuning parameters, $\lambda_1, \lambda_2, \lambda_3$, have been used in the second step, where $\lambda_1$ and $\lambda_2$ determine the sparsity of main and interaction effects with respect to the regulated and unregulated GEs correspondingly, and $\lambda_3$ controls the sparsity of the residuals from regulators. We choose the optimal tuning parameters using five-fold cross-validation in both the simulation study and real data analysis. The analysis has been implemented with statistical software R (version 3.6.3). In simulation, the average CPU time of running one replicated simulated data ($n = 500$, $p_g = p_r = 200$, $q = 4$) is 23.1 min. on a regular desktop PC. The R codes are available from the corresponding author.

## 3. Simulation

We perform simulation in order to evaluate the utility of the proposed method integrative G×E model, termed IGE. In addition, we consider three alternative methods: (1) the S-LASSO selects gene expressions and regulators separately using LASSO. (2) The J-LASSO selects gene expressions and regulators that are based on LASSO simultaneously. (3) ColReg, the collaborative regression [16], identifies important GEs and regulators jointly in terms of explaining similar variation under the cancer outcome.

We generate the data, as follows. First, each row of $R$ is independently generated from a multivariate normal distribution with mean zero and one of the four covariance structures: (i) AR–1 structure with correlation coefficient $0.25^{|i-j|}$ for the $i$th and $j$th regulators; (ii) banded correlation structure, where the $i$th and $j$th regulators have $\rho = 0.33$ if $|i-j| = 1$ and $\rho = 0$ otherwise; (iii) the covariance that was extracted from TCGA lung squamous cell carcinoma (LUSC) data in Section 4; and, (iv) the covariance structure of the lung adenocarcinoma (LUAD) from Section 4.

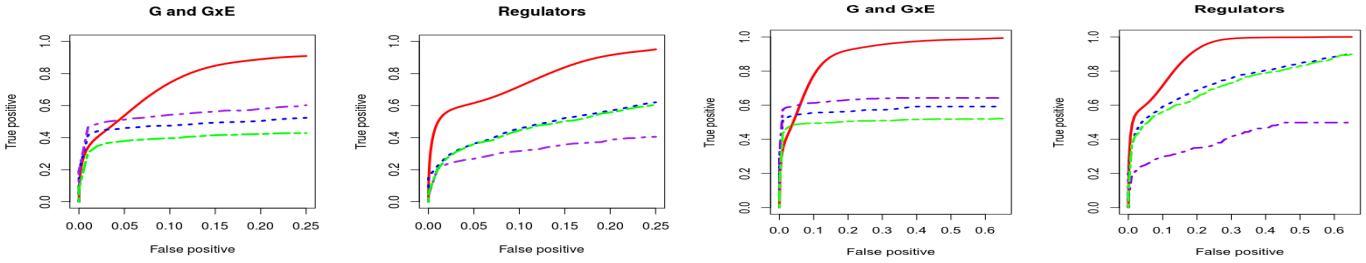
Choose $L = 20$ for the number of LRMs between gene expression and regulators. For $l = 1, \cdots, 20$, $u_l$ or $v_l$ is randomly assigned five non-zero entries, with values being generated from unif $[2, 4]$. Subsequently, $\Theta$ is computed as $\sum_{l=1}^{20} u_l v_l^\top$ and $G$ is generated as $G = R\Theta + \varepsilon$, where each row of matrix $\varepsilon$ is independently generated from a multivariate normal distribution with mean zero and the same covariance structure as $R$. To generate the cancer outcome, each row of $E$ is generated independently from a multivariate normal distribution with marginal mean zero and AR-1 structure, where the $i$th and $j$th components have correlation coefficient $0.5^{|i-j|}$. Subsequently, we generate the response from model (1) under standard normal errors.

200 gene expression, 200 regulators, and four environmental factors are simulated with two different sample sizes, 500 and 1000. We randomly select 30 gene expressions to assign non-zeros effects in model (1). For every selected gene expression, four non-zero entries are randomly assigned to the coefficients of G factor or its corresponding G×E interactions. Those values are generated from unif $[0.25, 0.5]$ and unif $[0.5, 1]$ for weak and strong coefficient signals, respectively. The coefficients of regulators are randomly assigned with 30 non-zero coefficients being generated from unif $[1, 2]$. The coefficients of environmental factors are generated from unif $[2, 3]$.

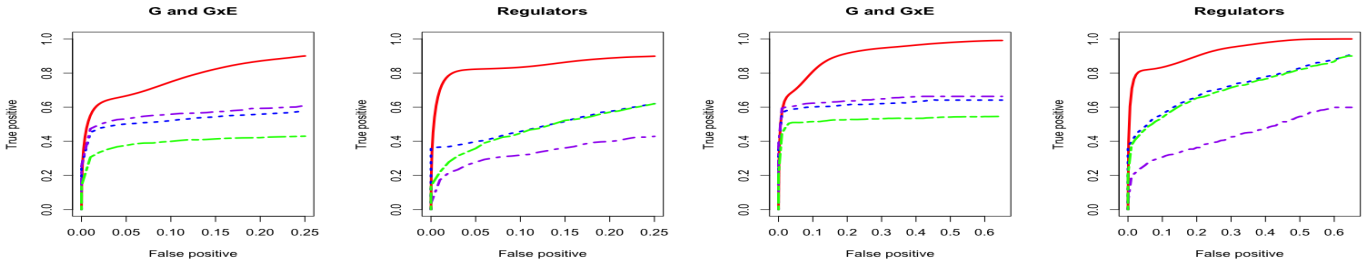
For a comprehensive evaluation, we consider a sequence of tuning parameter values (from 0 to 3, total 100 lambda values) and then use the receiver operating characteristic (ROC) curve and partial area under the ROC curve (PAUC) to compare the different methods. The total simulation replication is 100. All of the PAUCs are tabulated in Tables 1 and 2. Figures 1 and 2 show the ROC curves for the AR-1 structure and estimated covariance from LUSC. Appendix A provides other scenarios of ROC curves, respectively.

We consider using the receiver operating characteristic (ROC) curve and the partial area under the ROC curve (PAUC) to compare different methods. Total simulation replicates is 100. Tables 1 and 2 tabulate all of the PAUCs. Figures 1 and 2 show the ROC curves for AR-1 structure and estimated covariance from LUSC. Appendix A provides the ROC curves in other scenarios. For all simulation scenarios, the proposed method has higher PAUCs than the alternative methods. For example, in Table 1 with AR-1 correlation and weak signal, the proposed method has PAUC 0.73 (sd 0.07) for the identification of G and G×E effects, while J-LASSO, S-LASSO, and ColReg have PAUCs 0.54 (sd 0.04), 0.47 (sd 0.04), and 0.39 (sd 0.03), respectively. For the identification of regulators, the proposed method has PAUC 0.76 (sd 0.10), while J-LASSO, S-LASSO, and ColReg have PAUCs 0.32 (sd 0.05), 0.46 (sd 0.13), and 0.45 (sd 0.15), respectively. The similar pattern can be observed under settings with strong signals. When the sample size increases, the identification results of all methods become better. The proposed IGE outperforms alternative approaches across different scenarios. For instance, in Table 2 with AR-1 correlation and strong signal, the proposed method has PAUC 0.89 (sd 0.02) in the identification of G and G×E, while J-LASSO, S-LASSO, and ColReg have PAUCs 0.62 (sd 0.04), 0.57 (sd 0.04), and 0.50 (sd

0.03), correspondingly. For the identification of regulators, the proposed method also outperforms the alternatives.

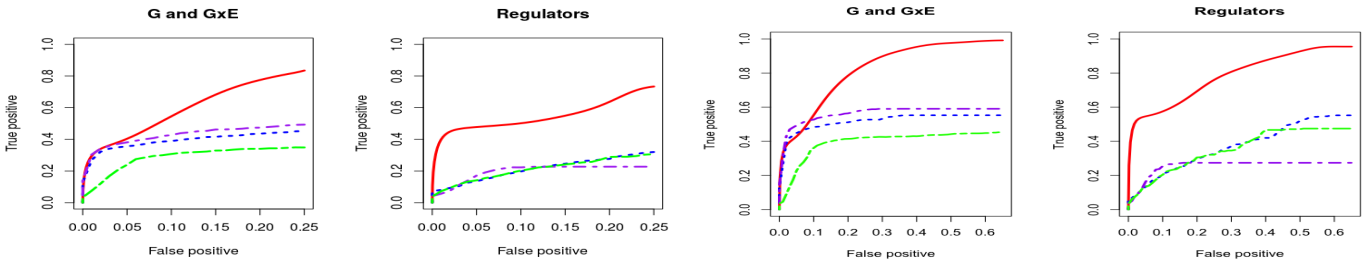



(**a**) n = 500, weak signal (**b**) n = 1000, weak signal

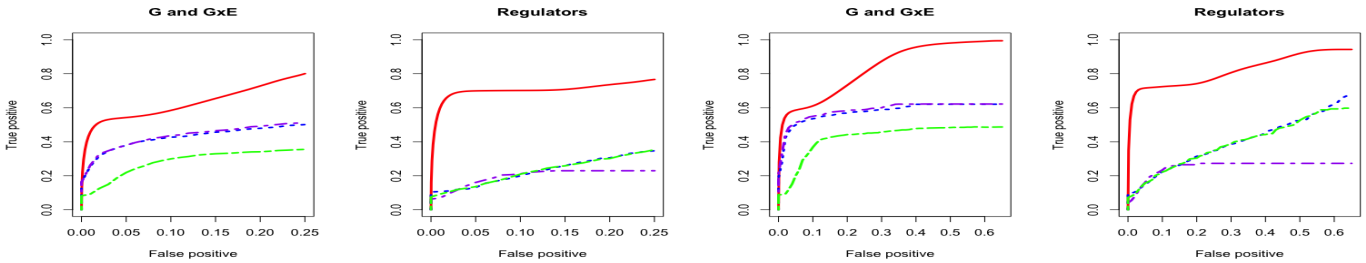



(**c**) n = 500, strong signal (**d**) n = 1000, strong signal

**Figure 1.** Four cases of receiver operating characteristic (ROC) curves under AR-1 structure. The left panel corresponds to comparison under both weak and strong signals for 500 subjects. The right panel corresponds to comparison under both weak and strong signals for 1000 subjects. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.




(**a**) n = 500, weak signal (**b**) n = 1000, weak signal




(**c**) n = 500, strong signal (**d**) n = 1000, strong signal

**Figure 2.** Four cases of ROC curves under estimated covariance from lung squamous cell carcinoma (LUSC). The left panel corresponds to comparison under both weak and strong signals for 500 subjects. The right panel corresponds to comparison under both weak and strong signals for 1000 subjects. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.

**Table 1.** PAUC: mean (sd) based on 100 replicates. $p_g = p_r = 200, n = 500$.

| Covariance | Signal | Approaches | G and G×E | Regulators |
|---|---|---|---|---|
| AR-1 | weak | IGE | 0.73 (0.07) | 0.76 (0.10) |
| | | S-LASSO | 0.47 (0.04) | 0.46 (0.13) |
| | | J-LASSO | 0.54 (0.04) | 0.32 (0.05) |
| | | ColReg | 0.39 (0.03) | 0.45 (0.15) |
| | strong | IGE | 0.77 (0.07) | 0.85 (0.06) |
| | | S-LASSO | 0.52 (0.05) | 0.48 (0.14) |
| | | J-LASSO | 0.55 (0.04) | 0.33 (0.05) |
| | | ColReg | 0.39 (0.03) | 0.46 (0.15) |
| Banded | weak | IGE | 0.74 (0.06) | 0.74 (0.10) |
| | | S-LASSO | 0.48 (0.03) | 0.44 (0.11) |
| | | J-LASSO | 0.54 (0.05) | 0.32 (0.04) |
| | | ColReg | 0.39 (0.03) | 0.43 (0.12) |
| | strong | IGE | 0.77 (0.08) | 0.84 (0.06) |
| | | S-LASSO | 0.52 (0.04) | 0.46 (0.11) |
| | | J-LASSO | 0.55 (0.05) | 0.32 (0.04) |
| | | ColReg | 0.39 (0.03) | 0.43 (0.12) |
| LUSC | weak | IGE | 0.59 (0.09) | 0.55 (0.15) |
| | | S-LASSO | 0.39 (0.04) | 0.21 (0.06) |
| | | J-LASSO | 0.42 (0.05) | 0.19 (0.06) |
| | | ColReg | 0.28 (0.04) | 0.21 (0.07) |
| | strong | IGE | 0.63 (0.10) | 0.71 (0.13) |
| | | S-LASSO | 0.42 (0.05) | 0.22 (0.07) |
| | | J-LASSO | 0.43 (0.05) | 0.19 (0.06) |
| | | ColReg | 0.28(0.05) | 0.22 (0.07) |
| LUAD | weak | IGE | 0.64 (0.09) | 0.62 (0.15) |
| | | S-LASSO | 0.45 (0.04) | 0.21 (0.06) |
| | | J-LASSO | 0.47 (0.05) | 0.19 (0.05) |
| | | ColReg | 0.32 (0.03) | 0.22 (0.07) |
| | strong | IGE | 0.70 (0.08) | 0.77 (0.11) |
| | | S-LASSO | 0.47 (0.05) | 0.23 (0.08) |
| | | J-LASSO | 0.48 (0.05) | 0.18 (0.05) |
| | | ColReg | 0.31 (0.04) | 0.23 (0.08) |

In addition, the proposed method outperforms the alternatives when the correlation is extracted from real data. For example, in Table 1, with estimated covariance from LUSC and weak signals, the proposed method has close PAUCs in both G and G×E and regulators, 0.59 (sd 0.09) and 0.55 (sd 0.15). Other methods have low accuracy in identifying main and interaction effects. In particular, J-LASSO, S-LASSO, and ColReg have PAUCs 0.42 (sd 0.05) and 0.19 (sd 0.06), 0.39 (sd 0.04) and 0.21 (sd 0.06), and 0.28 (sd 0.04) and 0.21 (sd 0.07), respectively. When magnitude of the signals and sample size increase (e.g., with LUSC and strong signals), the proposed method still have the best performance in identification. Overall, the IGE model has much higher identification accuracy than other methods across different simulation settings by borrowing strength from accounting for regulatory relationship and bi-level selection in G×E interaction studies.

**Table 2.** PAUC: mean (sd) based on 100 replicates. $p_g = p_r = 200, n = 1000$.

| Covariance | Signal | Approaches | G and G×E | Regulators |
|---|---|---|---|---|
| AR-1 | weak | IGE | 0.89 (0.02) | 0.91 (0.02) |
| | | S-LASSO | 0.57 (0.04) | 0.73 (0.09) |
| | | J-LASSO | 0.62 (0.04) | 0.40 (0.04) |
| | | ColReg | 0.50 (0.03) | 0.71 (0.09) |
| | strong | IGE | 0.91 (0.02) | 0.93 (0.02) |
| | | S-LASSO | 0.61 (0.04) | 0.71 (0.08) |
| | | J-LASSO | 0.64 (0.05) | 0.43 (0.04) |
| | | ColReg | 0.52 (0.03) | 0.70 (0.09) |
| Banded | weak | IGE | 0.89 (0.03) | 0.91 (0.03) |
| | | S-LASSO | 0.55 (0.04) | 0.73 (0.07) |
| | | J-LASSO | 0.62 (0.04) | 0.40 (0.05) |
| | | ColReg | 0.50 (0.03) | 0.71 (0.08) |
| | strong | IGE | 0.90 (0.04) | 0.92 (0.02) |
| | | S-LASSO | 0.61 (0.04) | 0.72 (0.08) |
| | | J-LASSO | 0.64 (0.04) | 0.44 (0.06) |
| | | ColReg | 0.53 (0.04) | 0.70 (0.08) |
| LUSC | weak | IGE | 0.82 (0.04) | 0.78 (0.06) |
| | | S-LASSO | 0.51 (0.05) | 0.36 (0.07) |
| | | J-LASSO | 0.56 (0.05) | 0.25 (0.07) |
| | | ColReg | 0.39 (0.04) | 0.35 (0.08) |
| | strong | IGE | 0.83 (0.04) | 0.82 (0.06) |
| | | S-LASSO | 0.57 (0.05) | 0.39 (0.07) |
| | | J-LASSO | 0.58 (0.05) | 0.25 (0.08) |
| | | ColReg | 0.42 (0.04) | 0.38 (0.07) |
| LUAD | weak | IGE | 0.83 (0.04) | 0.80 (0.06) |
| | | S-LASSO | 0.57 (0.04) | 0.43 (0.06) |
| | | J-LASSO | 0.59 (0.04) | 0.25 (0.06) |
| | | ColReg | 0.47 (0.03) | 0.43 (0.06) |
| | strong | IGE | 0.85 (0.03) | 0.84 (0.04) |
| | | S-LASSO | 0.61 (0.04) | 0.46 (0.07) |
| | | J-LASSO | 0.61 (0.04) | 0.26 (0.06) |
| | | ColReg | 0.49 (0.03) | 0.46 (0.07) |

## 4. Analysis of TCGA Data

Lung cancer is a top rank common cancer for both men and women. In this section, we apply the proposed method as well as the alternatives on lung adenocarcinoma (LUAD) data and lung squamous cell carcinoma (LUSC) data from the Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/).

At present, LUAD is the most common lung cancer subtype among non-smokers and women, although it has been shown that smoking may increase the risk of LUAD [17,18]. On the other hand, LUSC is closely associated with smoking, and it is more common in men than in women [19]. LUAD grows more slowly with smaller masses than LUSC of the same stage, but LUAD tends to initiate metastasis at the early stages [20].

The processed level 3 data have been downloaded from TCGA data portal while using package *cgdsr*. We match the multi-omics measurements with the clinical/environmental variables and survival outcome. LUSC and LUAD has 344 and 426 subjects, correspondingly. We first conduct screenings to reduce dimensionality, so the regularization methods can be appropriately applied. Here, we select the top 200 mRNA with the largest marginal variances. As we matched the CNA and Methylation profiles with same mRNA, the corresponding 200 measurements on CNA and Methylation are selected at the same time. We select age, gender, smoking pack years, and pathologic tumor stage as environmental variables. The accelerated failure time (AFT) model (Appendix B) has been adopted in order to link the omics and clinical measurements to survival outcomes.

### *4.1. Lung Adenocarcinoma (LUAD) Data*

The proposed method identifies eight LRMs with one residual effect of gene expression (mRNA) and 14 residual effects of regulators (DM and CNA). Additionally, the proposed method results in the identification of seven LRM×E interactions and 11 G×E interactions from mRNA residual effects.

Table 3 provides the identified main effects of LRMs, residual GEs, and regulators. We can observe that LRMs does not contain effects from methylation, while most of the residual effects in regulators are from methylation.The identification results have important biological implications. As a representative example, gene PIK3R2 is identified by 6 different LRMs. From a recent study [21], PIK3R2 is significantly associated with lung adenocarcinoma and its pathway plays a critical role in the progress of LUAD. Besides, gene STK3 is identified by five different LRMs. STK3 belongs to a large family of serine/threonine kinases, which are implicated in the regulation of signaling pathways involved in cell growth, differentiation and death. [22,23]. The identified LRMs are also meaningful. For example, we observe the regulatory relationship between PIK3R2 and NEK2 from both LRM #1 and #6. One of the recent studies shows that this natural downstream regulation is significantly related to cancer outcome [24]. Among all of the residual effects, we observe that most of them are from methylation. For example, SLC2A1, ECT2, TNS4, DKK1, and GNPNAT1 are found to be associated with the survival of lung cancer patients [25–29].

**Table 3.** Analysis of the the Cancer Genome Atlas (TCGA) lung adenocarcinoma (LUAD) data: linear regulatory models (LRMs) and residual effects for gene expression and regulators with the estimated coefficient or loadings in the parentheses.

| | **LRMs** | | | |
|---|---|---|---|---|
| | #1 (0.07) | #2 (−0.01) | #3 (−0.02) | #4 (−0.03) |
| mRNA | PIK3R2 (0.35)<br>STK3 (−0.74)<br>NCKAP5L (0.74)<br>CUL9 (0.14) | PIK3R2 (0.98)<br>STK3 (0.11)<br>NCKAP5L (−0.08) | ECT2 (−0.98)<br>PSMD2 (−0.17) | INTS7 (−0.77)<br>PIK3R2 (−0.62) |
| CNA | NEK2(−0.22)<br>LPGAT1 (0.22)<br>INTS7 (0.65)<br>DTL (−0.65)<br>CECR1 (−0.19) | CECR1 (0.65)<br>C1QTNF6 (−0.75) | KPNA4 (−0.44)<br>B3GALNT1 (0.43)<br>PSMD2 (−0.55)<br>LIPH (0.55) | INTS7 (−0.70)<br>DTL (0.70) |
| | #5 (−0.05) | #6 (0.08) | #7 (−0.06) | #8 (0.06) |
| mRNA | PIK3R2 (0.12)<br>STK3 (−0.78)<br>NCKAP5L (0.57)<br>CUL9 (0.16) | INTS7 (0.73)<br>PIK3R2 (0.63)<br>STK3 (0.18)<br>NCKAP5L (−0.14) | PIK3R2 (−0.10)<br>STK3 (−0.24)<br>CUL9 (−0.96) | PSMD2 (0.31)<br>TMOD 3(0.61)<br>DIAPH3 (0.72) |
| CNA | INTS7 (−0.16)<br>DTL (0.16)<br>CECR1 (−0.78)<br>C1QTNF6 (−0.57) | NEK2 (−0.69)<br>LPGAT1 (0.71) | INTS7 (−0.34)<br>DTL (0.36)<br>CECR1 (0.61)<br>C1QTNF6 (−0.61) | MAPRE3 (0.70)<br>IFT172 (−0.67)<br>PSMD2 (0.09)<br>ITGB1 (0.09)<br>ADAM10 (0.14) |
| | Residual effects | | | |
| mRNA | MAST3 (0.01) | | | |
| DM | ADSS (0.01)<br>TNS4 (0.02)<br>GNPNAT1 (0.04) | SLC2A1 (0.01)<br>MUSTN1 (0.05)<br>HPS1 (−0.04) | PTCH2 (0.01)<br>DKK1 (0.02)<br>MAPRE3 (−0.02) | ECT2 (0.09)<br>FSCN1 (0.05) |
| CNA | LAMC2 (−0.01) | CD5 (−0.03) | E2F7 (−0.01) | |

Table 4 provides the identification results for interaction effects. The proposed method selects variables with a sparse group nature. There are five LRMs interacting with environments. The first and fourth LRMs interact with two environment factors, and the

second, third, and fifth interact with one environment factor. Additionally, the proposed method can identify a total of 11 interactions involving mRNA residual effects. Note that, here, the G factor is no longer in the usual sense from existing G×E studies. The G factor are represented by the LRMs and residual mRNAs that correspond to the regulated and un-regulated G factors, respectively.

**Table 4.** Analysis of the TCGA LUAD data: G×E interaction identifications from LRMs and gene expression with the estimated regression coefficients in the parentheses.

| LRMs | AGE | GENDER | SMOKING |
|---|---|---|---|
| #1 | 0.08 | | −0.25 |
| #2 | | 0.02 | |
| #3 | | 0.01 | |
| #4 | | 0.01 | 0.01 |
| #5 | | | 0.01 |
| mRNA Residual | AGE | GENDER | SMOKING |
| MAST3 | | | 0.27 |
| HPS1 | 0.01 | | |
| BBS5 | −0.04 | | −0.03 |
| TLE1 | −0.01 | | |
| ADAM10 | | 0.02 | 0.03 |
| SLC16A3 | | 0.07 | |
| BTN2A2 | | −0.02 | −0.06 |
| FAM71E1 | | | 0.02 |

In terms of prediction, we adopt a random sampling approach. More specifically, we randomly select 30% data as a test set and the remaining as a training set. The estimates are generated using the training set only and the predictions are made based on the testing set. We dichotomize the predicted response at the median, create two risk groups, and compute log-rank statistics, which measure the difference in survival between the two groups. Larger log-rank test statistic indicates better predictive performance. The procedure is repeated 100 times to avoid extreme splits. The average log-rank test statistics are 5.97 (IGE, sd 0.35), 4.76 (S-LASSO, sd 0.25), 4.60 (J-LASSO, sd 0.08), and 3.74 (ColReg, sd 0.26), respectively. The proposed method has the largest log-rank statistic, hence the best prediction performance.

### *4.2. Lung Squamous Cell Carcinoma (LUSC) Data*

The proposed method identifies eight LRMs with two residual effects from GEs and 17 residual effects from regulators (DM and CNA). The interactions involve seven LRMs and 26 mRNAs.

Table 5 provides the identified main effects using the proposed method. As aforementioned, we aim to find a sparse relationship between gene expressions and regulators. Therefore, a small subset of regulators are related to genes and vice versa. Table 6 provides the identifications of G×E interaction effects. There's one LRM not interacting with any other environmental factors. The findings have important implications. For instance, gene RNF24 is identified by 2 different LRMs (#1, #2). RNF24 is a membrane protein, which interacts with TRPC protein [30]. A recent study shows that RNF24 acts as one of the important factors for the prognosis of carcinoma [31]. RNF24 is also shown to be correlated with the occurrence of esophageal adenocarcinoma [32]. For DM, RGP1 is identified by three different LRMs (#4, #6, #7). RGP1 belongs to the regulation of guanosine diphosphate (GDP) reaction exchange, and it acts as a prognostic factor in cancer, according to Anand (2020) [33]. For CNA, CD163L1 is identified by three different LRMs (#1, #4, #8), and it can be used as a significant biomarker of cancer [34]. The identified LRMs are also meaningful. For example, the regulatory relationship between NCOR2 and TCTN2 can be identified in LRM #7. This result has also been observed in a regulatory network analysis [35].

Among all of the residual effects, LRAT, PLEKHA6, ACOT7, KLK6, PLEKHB1, FGFRL1, and FPR2 are associated with prognosis of LUSC patients from existing studies [36–41].

**Table 5.** Analysis of the TCGA LUSC data: LRMs and residual effects for gene expression and regulators with the estimated coefficient or loadings in the parentheses.

| | **LRMs** | | | |
|---|---|---|---|---|
| | #1 (−0.01) | #2 (0.01) | #3 (0.01) | #4 (−0.02) |
| mRNA | RNF24 (−0.17) | SEC23B (0.23) | REEP3 (−0.76) | AP2A2 (−0.59) |
| | ESM1 (−0.53) | RNF24 (−0.97) | FUT11 (−0.64) | PNPLA6 (−0.37) |
| | RASAL2 (−0.39) | | | RFX1 (−0.55) |
| | LAMC1 (−0.34) | | | XRN2 (0.45) |
| | DLGAP4 (−0.63) | | | |
| DM | DCBLD1 (0.09) | TCF7L2 (0.22) | | RGP1 (−0.52) |
| | CHI3L1 (0.18) | | | NCOR2 (0.27) |
| CNA | CD163L1 (−0.16) | ENTPD6 (0.68) | RERE (−0.89) | CD163L1 (0.70) |
| | DLGAP4 (−0.96) | ABHD12 (−0.69) | DLGAP4 (−0.43) | PARD6G (−0.39) |
| | #5 (0.16) | #6 (0.05) | #7 (−0.05) | #8 (0.01) |
| mRNA | COL5A3 (0.45) | MGST3 (0.33) | TPM4 (0.68) | TCTN2 (−0.45) |
| | DCBLD1 (0.57) | OSBPL5 (0.31) | UBB (0.59) | ANGPT2 (−0.40) |
| | PDGFA (0.31) | SNX9 (0.56) | NCOR2 (−0.42) | UBE4B (−0.37) |
| | CHST15 (0.45) | MYO1C (0.46) | | MBTPS1 (−0.47) |
| | LGALS1 (0.39) | CCDC68 (0.49) | | FAM178B (−0.50) |
| DM | DCBLD1 (−0.86) | CHST15 (−0.97) | RGP1 (−0.55) | NCOR2 (0.16) |
| | FAM178B (−0.37) | RGP1 (0.13) | | |
| | CHST15 (−0.17) | NCOR2 (−0.10) | | |
| | | LGALS1 (−0.15) | | |
| CNA | DLGAP4 (0.27) | | STK40 (−0.26) | CD163L1 (−0.35) |
| | | | TCTN2 (−0.78) | DLGAP4 (−0.92) |
| | Residual effects | | | |
| mRNA | LRAT (−0.02) | PLEKHA6 (−0.02) | | |
| DM | BAMBI (0.01) | PYGB (0.02) | FUT11 (−0.18) | ZNF394 (0.03) |
| | CCIN (−0.01) | DEAF1 (−0.10) | ACOT7 (0.04) | KLK6 (−0.12) |
| | LHX8 (−0.01) | PLEKHB1 (0.09) | | |
| CNA | FGFRL1 (−0.05) | DCBLD1 (−0.04) | NEFL (−0.04) | CHST1 (0.02) |
| | ULK1 (−0.03) | FPR2 (0.02) | PYGB (−0.10) | |

We adopt a random sampling approach and apply log-rank test for assessment in order to evaluate prediction. We adopt the similar procedure as previous real data analysis section. After repeating 100 times, the average log-rank test statistics are 33.20 (IGE, sd 2.32), 25.06 (S-LASSO, sd 1.84), 24.41 (J-LASSO, sd 2.13), and 27.88 (ColReg, sd 2.45), respectively. The proposed method has superior prediction performance over alternatives.

**Table 6.** Analysis of the TCGA LUSC data: G×E interaction identifications from LRMs and gene expression with the estimated regression coefficients in the parentheses.

| **LRMs** | **AGE** | **GENDER** | **SMOKING** |
|---|---|---|---|
| #1 | | 0.02 | 0.03 |
| #2 | | 0.03 | |
| #4 | −0.02 | | |
| #5 | 0.01 | 0.05 | −0.02 |
| #6 | 0.01 | −0.01 | |
| #7 | | −0.36 | |
| #8 | | 0.02 | |
| mRNA Residual | AGE | GENDER | SMOKING |
| LRAT | | −0.17 | |
| PLEKHA6 | | −0.30 | |
| AP2A2 | 0.02 | | |
| SLC12A7 | −0.10 | 0.07 | |
| TCTN2 | −0.15 | −0.09 | |
| CLEC5A | 0.01 | | |
| RNF24 | −0.06 | 0.04 | |
| PRRX2 | 0.04 | | −0.04 |
| CCDC74A | 0.14 | −0.13 | |
| FGF9 | 0.03 | | −0.06 |
| IGF2R | 0.05 | −0.02 | |
| CHMP4C | 0.24 | 0.13 | −0.01 |
| SLC45A4 | −0.11 | | |
| SULF2 | −0.05 | −0.03 | |
| UBB | | −0.11 | |
| DVL1 | | −0.07 | |
| NID1 | | 0.08 | 0.20 |
| KLK8 | | 0.01 | |
| DOCK6 | | 0.26 | −0.10 |
| FHDC1 | | 0.01 | −0.16 |
| OPLAH | | −0.12 | |
| VSTM1 | | | −0.02 |
| SLC28A1 | | | −0.07 |
| TCF7L2 | | | 0.12 |
| DLGAP4 | | | −0.04 |
| CRNKL1 | | | −0.25 |

## 5. Discussion

We have conducted an integrative gene–environment interaction analysis for multi-dimensional omics data based on the proposed two-step variable selection model. Specifically, at the first step, sparse regulatory relationship between the G factor and its regulators have been pinpointed via penalization, which leads to effects that can be directly linked to the prognostic outcomes. At the second step, a G×E prognostic model has been considered, where the G factor that is involved in the interaction consists of regulated (corresponding to the LRM) and unregulated (i.e., the residual GE) components. Besides, the residuals of the regulator are also included. The integrative G×E analysis fully takes the advantage of the multi-omics measurements, which distinguishes itself from most of the published studies.

Traditionally, statistical testing based marginal analysis has dominated the G×E studies. The paradigm shift to the joint analysis has been mainly motivated by the gene set and pathway-based association analysis [42–45]. Recently, the effectiveness of regularized variable selection has been recognized not only in joint G×E studies when a large number of genetic factors are involved [7], but also in multi-level omics integrations [5]. Therefore, it has been adopted here.

This study can be improved by the following aspects. Because strong correlations have been widely observed in among omics measurements, network based penalization can be imposed to accommodate the correlations among regulators at the first stage [46–48].

Besides, robustness can be incorporated at the first stage to model the regulatory relationship between GE and its regulators [49], and in the second stage for a robust prognostic model [50,51]. Accounting for the form of environmental factors has received considerable attention in G×E studies, which results in the development of a wide range of nonparametric [52–54] and semiparametric [55–57] methods. However, in integrative G×E studies, capturing the nonlinear form of interaction is challenging. In this study, we focus on prognostic outcomes. With other types of outcomes, such as the longitudinal phenotypes [58,59], the G×E model in the second stage can be modified accordingly.

**Author Contributions:** Conceptualization, Y.D. and C.W.; methodology, Y.D., K.F., X.L. and C.W.; software, Y.D. and C.W.; formal analysis, Y.D.; investigation, Y.D.; writing—original draft preparation, Y.D. and C.W.; writing—review and editing, Y.D., K.F., X.L. and C.W.; visualization, Y.D.; supervision, C.W.; funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study received no external funding. It has been partly supported by an innovative research award from KSU Johnson Cancer Research Center and a KSU Faculty Enhancement Award.

**Institutional Review Board Statement:** This study is a secondary data analysis. The dataset can be freely downloaded through TCGA data portal. The IRB is not required for accessing and using the data. The patient information has been de-identified from the dataset used in this study.

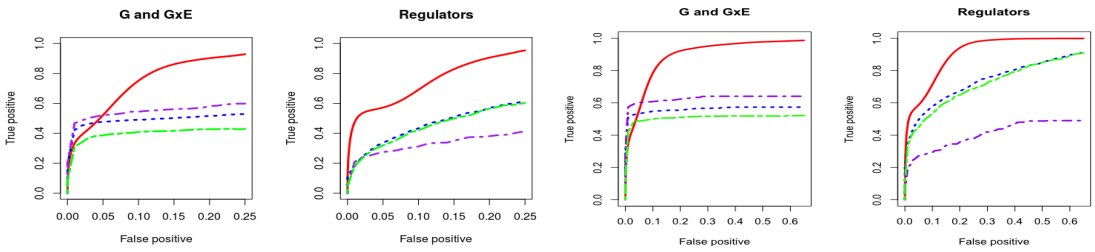
**Informed Consent Statement:** Not applicable due to the reason specified above.

**Data Availability Statement:** The datasets used for the analyses described in this manuscript has been downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/) and are available to the general public without restricted access.

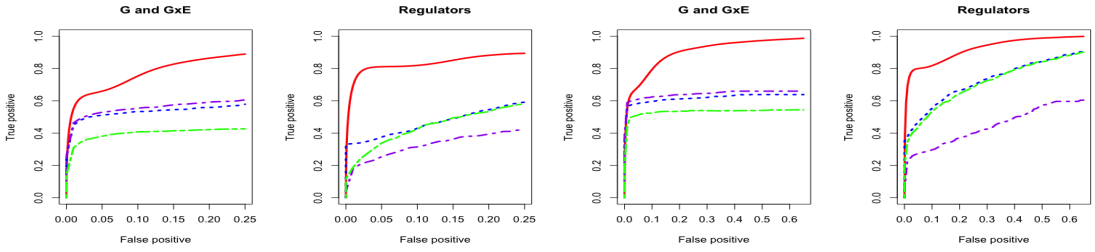
**Acknowledgments:** We thank the editor and reviewers for their invitation, careful review and insightful comments, leading to a significant improvement of this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Other Simulation Scenarios

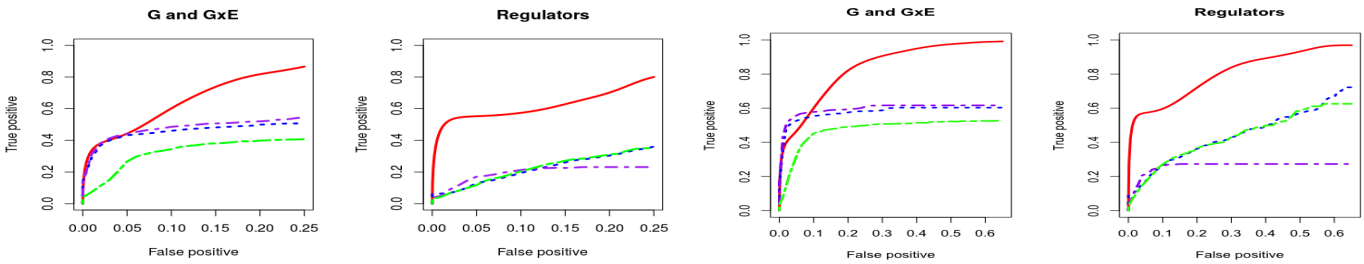



(**a**) n = 500, weak signal (**b**) n = 1000, weak signal




(**c**) n = 500, strong signal (**d**) n = 1000, strong signal

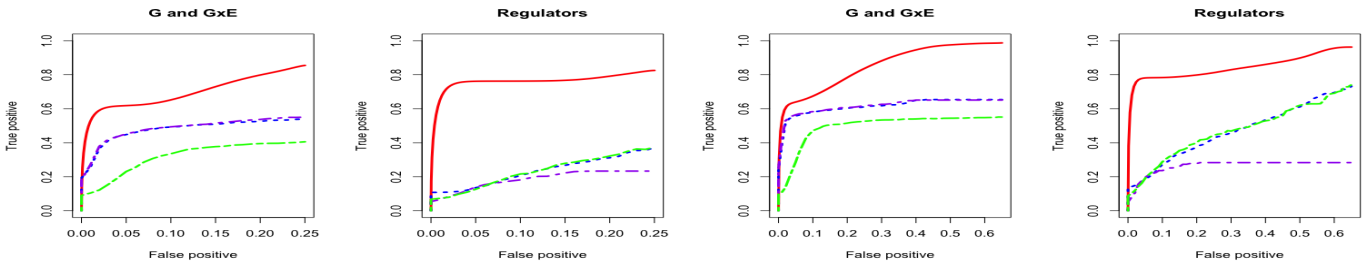
**Figure A1.** Four cases of ROC curves under banded correlation structure. Left two columns are 500 subjects to compare weak and strong signal performance. Right two columns are 1000 subjects to compare weak and strong signal performance. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.




(**a**) n = 500, weak signal




(**b**) n = 1000, weak signal




(**c**) n = 500, strong signal




(**d**) n = 1000, strong signal

**Figure A2.** Four cases of ROC curves under estimated covariance from LUAD. Left two columns are 500 subjects to compare weak and strong signal performance. Right two columns are 1000 subjects to compare weak and strong signal performance. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.

## Appendix B. Accelerated Failure Time (AFT) Model

Denote $T$ as the logarithm of the failure time and denote $C$ as the logarithm of the censoring time. Under right censoring, we observe $Y = \min(T, C)$, $\delta = I(T \leq C)$. We adopt the Kaplan-Meier weights for censoring. Let $\hat{F}$ be the Kaplan-Meier estimator of the distribution function $F$ of $T$. According to [60], we have $\hat{F}(y) = \sum_{i=1}^{n} w_i I\{Y_{(i)} \leq y\}$, where $w_i$ can be computed as

$$w_1 = \frac{\delta_{(1)}}{n}, w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left( \frac{n-j}{n-j+1} \right)^{\delta_j}, \; i = 2, ..., n,$$

where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of $Y_i$ and $\delta_{(1)}, \cdots, \delta_{(n)}$ are the corresponding censoring indicators. Denote $(E_{(i)}, X_{1_{(i)}}, X_{2_{(i)}}, \tilde{R}_{(i)})$ as the measurements associated with $(Y_{(i)}, \delta_{(i)})$, where the notations are from Equation (9). We center $E_{(i)}, X_{1_{(i)}}, X_{2_{(i)}}, \tilde{R}_{(i)}, Y_{(i)}$ using $w_i$-weighted mean as follows:

$$\bar{E}_w = \sum_{i=1}^{n} w_i E_{(i)} / \sum_{i=1}^{n} w_i, \; \bar{X}_{1_w} = \sum_{i=1}^{n} w_i X_{1_{(i)}} / \sum_{i=1}^{n} w_i, \bar{X}_{2_w} = \sum_{i=1}^{n} w_i X_{2_{(i)}} / \sum_{i=1}^{n} w_i,$$

$$\bar{\tilde{R}}_w = \sum_{i=1}^{n} w_i \tilde{R}_{(i)} / \sum_{i=1}^{n} w_i, \bar{Y}_w = \sum_{i=1}^{n} w_i Y_{(i)} / \sum_{i=1}^{n} w_i.$$

Then the centered predictors and responses are $E_{w_{(i)}} = \sqrt{w_i}(E_{(i)} - \bar{E}_w)$, $X_{1_{w_{(i)}}} = \sqrt{w_i}(X_{1_{(i)}} - \bar{X}_{1_w})$, $X_{2_{w_{(i)}}} = \sqrt{w_i}(X_{2_{(i)}} - \bar{X}_{2_w})$, $\tilde{R}_{w_{(i)}} = \sqrt{w_i}(\tilde{R}_{(i)} - \bar{\tilde{R}}_w)$ and $Y_{w_{(i)}} = \sqrt{w_i}(Y_{(i)} - \bar{Y}_w)$. Hence, $Y = (Y_{w(1)}, \cdots, Y_{w(n)})^T$, $E = (E_{w(1)}, \cdots, E_{w(n)})^T$, $X_1 = (X_{1_{w(1)}}, \cdots, X_{1_{w(n)}})^T$, $X_2 = (X_{2_{w(1)}}, \cdots, X_{2_{w(n)}})^T$, and $\tilde{R} = (\tilde{R}_{w(1)}, \cdots, \tilde{R}_{w(n)})^T$.

## References


1. Simonds, N.I.; Ghazarian, A.A.; Pimentel, C.B.; Schully, S.D.; Ellison, G.L.; Gillanders, E.M.; Mechanic, L.E. Review of the gene-environment interaction literature in cancer: What do we know? *Genet. Epidemiol.* **2016**, *40*, 356–365. [CrossRef]
2. Dempfle, A.; Scherag, A.; Hein, R.; Beckmann, L.; Chang-Claude, J.; Schäfer, H. Gene-environment interactions for complex traits: Definitions, methodological requirements and challenges. *Eur. J. Hum. Genet.* **2008**, *16*, 1164–1172. [CrossRef] [PubMed]
3. Hirschhorn, J.N.; Lohmueller, K.; Byrne, E.; Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **2002**, *4*, 45–61. [PubMed]
4. Wu, C.; Li, S.; Cui, Y. Genetic association studies: An information content perspective. *Curr. Genom.* **2012**, *13*, 566–573. [CrossRef] [PubMed]
5. Wu, C.; Zhou, F.; Ren, J.; Li, X.; Jiang, Y.; Ma, S. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High-throughput* **2019**, *8*, 4.
6. Zhu, R.; Zhao, Q.; Zhao, H.; Ma, S. Integrating multidimensional omics data for cancer outcome. *Biostatistics* **2016**, *17*, 605–618. [CrossRef]
7. Zhou, F.; Ren, J.; Lu, X.; Ma, S.; Wu, C. Gene-Environment Interaction: A Variable Selection Perspective. *Epistasis Methods Mol. Biol.* **2021**, in press.
8. Wang, W.; Baladandayuthapani, V.; Morris, J.S.; Broom, B.M.; Manyam, G.; Do, K.A. iBAG: Integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **2013**, *29*, 149–159. [CrossRef]
9. Ciriello, G.; Cerami, E.; Sander, C.; Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **2012**, *22*, 398–406. [CrossRef]
10. Kristensen, V.N.; Lingjærde, O.C.; Russnes, H.G.; Vollan, H.K.M.; Frigessi, A.; Børresen-Dale, A.L. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* **2014**, *14*, 299. [CrossRef]
11. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
12. Lee, M.; Shen, H.; Huang, J.Z.; Marron, J. Biclustering via sparse singular value decomposition. *Biometrics* **2010**, *66*, 1087–1095. [CrossRef] [PubMed]
13. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
14. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360.
15. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
16. Gross, S.M.; Tibshirani, R. Collaborative regression. *Biostatistics* **2014**, *16*, 326–338.
17. Subramanian, J.; Govindan, R. Lung cancer in never smokers: A review. *J. Clin. Oncol.* **2007**, *25*, 561–570. [CrossRef]
18. Couraud, S.; Zalcman, G.; Milleron, B.; Morin, F.; Souquet, P.J. Lung cancer in never smokers—A review. *Eur. J. Cancer* **2012**, *48*, 1299–1311.
19. Kenfield, S.A.; Wei, E.K.; Stampfer, M.J.; Rosner, B.A.; Colditz, G.A. Comparison of aspects of smoking among the four histological types of lung cancer. *Tob. Control* **2008**, *17*, 198–204. [CrossRef]
20. Kumar, V.; Abbas, A.K.; Aster, J.C. *Robbins Basic Pathology e-book*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2017.
21. Chen, Y.; Tang, J.; Lu, T.; Liu, F. CAPN1 promotes malignant behavior and erlotinib resistance mediated by phosphorylation of c-Met and PIK3R2 via degrading PTPN1 in lung adenocarcinoma. *Thorac. Cancer* **2020**, *11*, 1848–1860. [CrossRef]
22. Huang, N.; Lin, W.; Shi, X.; Tao, T. STK24 expression is modulated by DNA copy number/methylation in lung adenocarcinoma and predicts poor survival. *Future Oncol.* **2018**, *14*, 2253–2263. [CrossRef] [PubMed]
23. Pombo, C.M.; Force, T.; Kyriakis, J.; Nogueira, E.; Fidalgo, M.; Zalvide, J. The GCK II and III subfamilies of the STE20 group kinases. *Front Biosci* **2007**, *12*, 850–859. [CrossRef] [PubMed]
24. Hameed, Y.; Ejaz, S. Up-regulation of FN1, Activation of Maturation Promoting Factor and Associated Signaling Pathway Facilitates Epithelial-Mesenchymal Transition, Inhibits Apoptosis and Elevates Proliferation Rate of Breast Cancer Cells. *Silico Anal. Microarray Datasets* **2020**. [CrossRef]
25. Guo, W.; Sun, S.; Guo, L.; Song, P.; Xue, X.; Zhang, H.; Zhang, G.; Li, R.; Gao, Y.; Qiu, B.; et al. Elevated SLC2A1 Expression Correlates with Poor Prognosis in Patients with Surgically Resected Lung Adenocarcinoma: A Study Based on Immunohistochemical Analysis and Bioinformatics. *DNA Cell Biol.* **2020**, *39*, 631–644.
26. Silva, V.M.; Gomes, J.A.; Tenório, L.P.G.; de Omena Neta, G.C.; da Costa Paixão, K.; Duarte, A.K.F.; da Silva, G.C.B.; Ferreira, R.J.S.; Koike, B.D.V.; de Sales Marques, C.; et al. Schwann cell reprogramming and lung cancer progression: A meta-analysis of transcriptome data. *Oncotarget* **2019**, *10*, 7288. [CrossRef]
27. Misono, S.; Seki, N.; Mizuno, K.; Yamada, Y.; Uchida, A.; Sanada, H.; Moriya, S.; Kikkawa, N.; Kumamoto, T.; Suetsugu, T.; et al. Molecular pathogenesis of gene regulation by the miR-150 duplex: miR-150-3p regulates TNS4 in lung adenocarcinoma. *Cancers* **2019**, *11*, 601. [CrossRef]
28. Yang, J.; Liu, Y.; Mai, X.; Lu, S.; Jin, L.; Tai, X. STAT1-induced upregulation of LINC00467 promotes the proliferation migration of lung adenocarcinoma cells by epigenetically silencing DKK1 to activate Wnt/$\beta$-catenin signaling pathway. *Biochem. Biophys. Res. Commun.* **2019**, *514*, 118–126. [CrossRef]
29. Zhang, S.; Lu, Y.; Liu, Z.; Li, X.; Wang, Z.; Cai, Z. Identification Six Metabolic Genes as Potential Biomarkers for Lung Adenocarcinoma. *J. Comput. Biol.* **2020**, *27*, 1532–1543. [CrossRef]

30. Lussier, M.P.; Lepage, P.K.; Bousquet, S.M.; Boulay, G. RNF24, a new TRPC interacting protein, causes the intracellular retention of TRPC. *Cell Calcium* **2008**, *43*, 432–443. [CrossRef]
31. Lin, T.; Gu, J.; Qu, K.; Zhang, X.; Ma, X.; Miao, R.; Xiang, X.; Fu, Y.; Niu, W.; She, J.; et al. A new risk score based on twelve hepatocellular carcinoma-specific gene expression can predict the patients' prognosis. *Aging (Albany N. Y.)* **2018**, *10*, 2480. [CrossRef]
32. Wang, X.W.; Wei, W.; Wang, W.Q.; Zhao, X.Y.; Guo, H.; Fang, D.C. RING finger proteins are involved in the progression of barrett esophagus to esophageal adenocarcinoma: A preliminary study. *Gut Liver* **2014**, *8*, 487. [CrossRef] [PubMed]
33. Anand, S.; Khan, M.A.; Khushman, M.; Dasgupta, S.; Singh, S.; Singh, A.P. Comprehensive Analysis of Expression, Clinicopathological Association and Potential Prognostic Significance of RABs in Pancreatic Cancer. *Int. J. Mol. Sci.* **2020**, *21*, 5580.
34. Zahra, A.; Rubab, I.; Malik, S.; Khan, A.; Khan, M.J.; Fatmi, M.Q. Meta-Analysis of miRNAs and their involvement as biomarkers in oral cancers. *BioMed Res. Int.* **2018**, *2018*, 8439820. [CrossRef] [PubMed]
35. Zeng, L.; Yu, J.; Huang, T.; Jia, H.; Dong, Q.; He, F.; Yuan, W.; Qin, L.; Li, Y.; Xie, L. Differential combinatorial regulatory network analysis related to venous metastasis of hepatocellular carcinoma. *BMC Genom.* **2012**, *13*, S14. [CrossRef]
36. Ke, D.; Guo, Q.; Fan, T.Y.; Xiao, X. Analysis of the Role and Regulation Mechanism of hsa-miR-147b in Lung Squamous Cell Carcinoma Based on The Cancer Genome Atlas Database. *Cancer Biother. Radiopharm.* **2020**.
37. Relli, V.; Trerotola, M.; Guerra, E.; Alberti, S. Abandoning the notion of non-small cell lung cancer. *Trends Mol. Med.* **2019**, *25*, 585–594. [CrossRef]
38. Zhang, Q.; Huang, R.; Hu, H.; Yu, L.; Tang, Q.; Tao, Y.; Liu, Z.; Li, J.; Wang, G. Integrative analysis of hypoxia-associated signature in pan-cancer. *iScience* **2020**, *23*, 101460. [CrossRef]
39. Wang, Y.; Zhang, J.; Xiao, X.; Liu, H.; Wang, F.; Li, S.; Wen, Y.; Wei, Y.; Su, J.; Zhang, Y.; et al. The identification of age-associated cancer markers by an integrative analysis of dynamic DNA methylation changes. *Sci. Rep.* **2016**, *6*, 22722.
40. Bae, J.M.; Wen, X.; Kim, T.S.; Kwak, Y.; Cho, N.Y.; Lee, H.S.; Kang, G.H. Fibroblast growth factor receptor 1 (FGFR1) amplification detected by droplet digital polymerase chain reaction (ddPCR) is a prognostic factor in colorectal cancers. *Cancer Res. Treat. Off. J. Korean Cancer Assoc.* **2020**, *52*, 74. [CrossRef]
41. Hu, J.; Xu, L.; Shou, T.; Chen, Q. Systematic analysis identifies three-lncRNA signature as a potentially prognostic biomarker for lung squamous cell carcinoma using bioinformatics strategy. *Transl. Lung Cancer Res.* **2019**, *8*, 614. [CrossRef]
42. Wang, L.; Jia, P.; Wolfinger, R.D.; Chen, X.; Zhao, Z. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics* **2011**, *98*, 1–8. [CrossRef] [PubMed]
43. Wu, C.; Cui, Y. Boosting signals in gene-based association studies via efficient SNP selection. *Briefings Bioinform.* **2014**, *15*, 279–291. [CrossRef] [PubMed]
44. Jin, L.; Zuo, X.Y.; Su, W.Y.; Zhao, X.L.; Yuan, M.Q.; Han, L.Z.; Zhao, X.; Chen, Y.D.; Rao, S.Q. Pathway-based analysis tools for complex diseases: A review. *Genom. Proteom. Bioinform.* **2014**, *12*, 210–220. [CrossRef] [PubMed]
45. Jiang, Y.; Huang, Y.; Du, Y.; Zhao, Y.; Ren, J.; Ma, S.; Wu, C. Identification of prognostic genes and pathways in lung adenocarcinoma using a Bayesian approach. *Cancer Inform.* **2017**, *16*, 1176935116684825. [CrossRef] [PubMed]
46. Li, C.; Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **2008**, *24*, 1175–1182. [CrossRef] [PubMed]
47. Sun, H.; Wang, S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics* **2012**, *28*, 1368–1375. [CrossRef] [PubMed]
48. Ren, J.; He, T.; Li, Y.; Liu, S.; Du, Y.; Jiang, Y.; Wu, C. Network-based regularization for high dimensional SNP data in the case–control study of Type 2 diabetes. *BMC Genet.* **2017**, *18*, 44. [CrossRef]
49. Wu, C.; Zhang, Q.; Jiang, Y.; Ma, S. Robust network-based analysis of the associations between (epi) genetic measurements. *J. Multivar. Anal.* **2018**, *168*, 119–130. [CrossRef]
50. Ren, J.; Du, Y.; Li, S.; Ma, S.; Jiang, Y.; Wu, C. Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genet. Epidemiol.* **2019**, *43*, 276–291.
51. Wu, C.; Jiang, Y.; Ren, J.; Cui, Y.; Ma, S. Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Stat. Med.* **2018**, *37*, 437–456. [CrossRef]
52. Li, J.; Wang, Z.; Li, R.; Wu, R. Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat.* **2015**, *9*, 640. [CrossRef] [PubMed]
53. Wu, C.; Cui, Y. A novel method for identifying nonlinear gene–environment interactions in case–control association studies. *Hum. Genet.* **2013**, *132*, 1413–1425. [CrossRef] [PubMed]
54. Wu, C.; Zhong, P.S.; Cui, Y. Additive varying-coefficient model for nonlinear gene-environment interactions. *Stat. Appl. Genet. Mol. Biol.* **2018**, *17*.
55. Wu, C.; Shi, X.; Cui, Y.; Ma, S. A penalized robust semiparametric approach for gene–environment interactions. *Stat. Med.* **2015**, *34*, 4016–4030. [CrossRef] [PubMed]
56. Ma, S.; Xu, S. Semiparametric nonlinear regression for detecting gene and environment interactions. *J. Stat. Plan. Inference* **2015**, *156*, 31–47. [CrossRef]
57. Ren, J.; Zhou, F.; Li, X.; Chen, Q.; Zhang, H.; Ma, S.; Jiang, Y.; Wu, C. Semiparametric Bayesian variable selection for gene-environment interactions. *Stat. Med.* **2020**, *39*, 617–638. [CrossRef]

58. Li, J.; Lu, Q.; Wen, Y. Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data. *Bioinformatics* **2020**, *36*, 1785–1794. [CrossRef]
59. Zhou, F.; Ren, J.; Li, G.; Jiang, Y.; Li, X.; Wang, W.; Wu, C. Penalized Variable Selection for Lipid–Environment interactions in a longitudinal lipidomics study. *Genes* **2019**, *10*, 1002. [CrossRef]
60. Stute, W.; Wang, J.L. The strong law under random censorship. *Ann. Stat.* **1993**, *21*, 1591–1607. [CrossRef]