Overlapping Group Screening for Detection of Gene-environment Interactions with Application to TCGA High-dimensional Survival Genomic Data

Jie-Huei Wang (✉ jhwang@mail.fcu.edu.tw )
Feng Chia University
Kang-Hsin Wang
Feng Chia University
Yi-Hau Chen
Academia Sinica

Research Article

Keywords: gene-environment interaction, joint model, Lasso, overlapping group screening, survival prediction, TCGA

Posted Date: January 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1283374/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Overlapping Group Screening for Detection of Gene-environment Interactions

with Application to TCGA High-dimensional Survival Genomic Data

Jie-Huei Wang*, Kang-Hsin Wang 1 and Yi-Hau Chen 2

jhwang@mail.fcu.edu.tw, kkwang92066@gmail.com, and yhchen@stat.sinica.edu.tw

Department of Statistics, Feng Chia University, Seatwen, Taichung 40724, Taiwan 1 and Institute of Statistical Science, Academia Sinica, Nankang, Taipei 11529, Taiwan2

Abstract

Background: In the context of biomedical and epidemiological research, gene-environment (G-E) interaction is of great significance to the etiology and progression of many complex diseases. In high-dimensional genetic data, two general models, marginal and joint models, are proposed to identify important interaction factors. Most existing approaches for identifying G-E interactions are limited owing to the lack of robustness to outliers/contamination in response and predictor data. In particular, right-censored survival outcomes make the associated feature screening even challenging. In this article, we utilize the overlapping group screening (OGS) approach to select important G-E interactions related to clinical survival outcomes by incorporating the gene pathway information under a joint modeling framework.

Results: Simulation studies under various scenarios are carried out to compare the performances of our proposed method with some commonly used methods. In the real data
applications, we use our proposed method to identify G-E interactions related to the clinical survival outcomes of patients with head and neck squamous cell carcinoma, esophageal carcinoma, and lung adenocarcinoma in The Cancer Genome Atlas clinical survival genetic data, and further establish corresponding survival prediction models. Both simulation and real data studies show that our method performs well and outperforms existing methods in the G-E interaction selection, effect estimation, and survival prediction accuracy.

**Conclusions:** The OGS approach is useful for selecting important environmental factors, genes and G-E interactions in the ultra-high dimensional feature space. The prediction ability of OGS with the Lasso penalty is better than existing methods. The same idea of the OGS approach can apply to other outcome models, such as the proportional odds survival time model, the logistic regression model for binary outcomes, and the multinomial logistic regression model for multi-class outcomes.

**Keywords:** gene-environment interaction, joint model, Lasso, overlapping group screening, survival prediction, TCGA

**Background**

It is believed that in the development of complex diseases such as cancer, diabetes, and so on, gene-environment (G-E) interaction plays a critical role beyond the main genetic (G) or environmental (E) factors ([1-2] and so on). For example, Batchelor et al. [3] showed that the
interaction between the gene TP53 and age affects the prognosis of glioblastoma. As a consequence, incorporating significant G-E interaction factors into a survival prediction model would enhance the performance of the later.

In the setting of high-dimensional genetic data analysis, there exist two ways to identification of important G-E interactions: the marginal and joint analyses [4]. The marginal analysis considers only one gene at a time, and fits a model consisting of multiple E factors, this gene, and its interaction with E factors. The other performs joint analysis and considers all genes in a single model.

In the framework of marginal analysis of high-dimensional genetic data, for each gene, a model consisting of multiple E factors, a single gene itself, and its interaction with E factors is fitted. Specifically, the conceptual marginal model is "Outcome~Es+G+G*(Es)", where the outcome variable can be a continuous, categorical, or survival time phenotype, Es represents a set of environmental factors such as environmental exposures, demographic, clinical, and socioeconomic variables, and G*(Es) represents the interaction between the G factor and all E factors under consideration. The significant G-E interactions can be selected based on the corresponding marginal p-values. Since the marginal model is low-dimensional, its main advantage is its computational stability and conceptual simplicity. Therefore, marginal programs are popular in the fields of bioinformatics and biomedicine. However, a common limitation of traditional methods of marginal analysis is its lack of robustness. In practical genetic studies, Xu et al. [5] pointed out that long-tailed distributions and contamination in
prognosis response and predictors are not uncommon. In addition, human input errors may also lead to long-tailed distributions and contamination. In Fig. 1, we analyzed The Cancer Genome Atlas (TCGA) clinical survival data for esophageal carcinoma (ESCA) to show the long-tailed distribution phenomenon. The long-tailed distributions of TCGA clinical survival data for head and neck squamous cell carcinoma (HNSCC) and lung adenocarcinoma (LUAD) are displayed in Appendix S.1. Moreover, censored survival outcomes make the relevant feature screening difficult.

![Fig. 1 The long-tailed distribution of clinical survival data for esophageal carcinoma (the TCGA ESCA)](image)

On the other hand, models in the framework of joint analysis better describe disease biology given the fact that complex diseases are related to the combined effects of multiple genetic biomarkers. The conceptual joint model is "Outcome~Es+Gs+(Gs)*(Es)", where Gs represents a set of G factors, including gene expressions, SNPs and other types of molecular
measurements, and \((Gs)\)*(Es) represents the interactions between all G and E factors. In this article, we focus on the joint analysis framework. A common challenge of joint analysis is its high dimensionality, which makes it difficult to identify significant interaction effects. Moreover, right-censored survival outcomes and contaminated biomarker data make the task even challenging.

For survival outcomes, popular models include the accelerated failure time (AFT) model and Cox’s model. Based on the AFT model, several robust joint regression methods have been proposed. The Penalized trimmed regression (PTReg) method [6] uses the trimmed regression to account for long-tailed distribution/contamination in prognosis response and predictors, and Wu et al. [7] incorporates the G structure into the joint modeling. These methods conduct regularized estimation and selection based on the minimax concave penalty (MCP) penalty and utilize a decomposition technique to explain the interaction hierarchy. Their main potential disadvantage is that the model size is much larger than the sample size, and the statistical power under the penalized regression frameworks may be suboptimal [8]. In addition, since the gene expression data is often contaminated, the traditional Pearson correlation or Gaussian graphical models may not be a suitable measure to quantify the correlation among genes [9].

Based on the above rationale, we plan to adopt a two-step screening approach to detect G-E interactions by incorporating biological pathways information. The proposed method uses annotated gene sets collected in the molecular signatures database [10], which can be downloaded from the website [http://www.broadinstitute.org/gsea/msigdb](http://www.broadinstitute.org/gsea/msigdb). Wang and Chen [11]
described the idea of an overlapping group screening procedure, called the OGS method, for survival prediction based on the Cox model. In this work, we extend the OGS method to detect G-E interactions, and show that OGS has several advantages: (i) it can alleviate the collinearity problem in regression analysis due to the correlation between biomarkers in the same gene/pathway; (ii) it can significantly reduce the search space for interaction effects by using the feature grouping structure; and (iii) it can significantly improve the model selection performed by penalized regression in an ultra-high dimensional feature space.

Simulation studies under various scenarios reveal that our method works well and outperforms existing methods in the model selection, estimation, and prediction accuracy. In the real data application, we combine gene expression profile data with prior pathway information from the Gene Ontology biological process (GO-BP) database and use the OGS approach to select several important environmental factors, genes, and G-E interactions that are associated with clinical survival outcomes of patients with HNSCC, ESCA, and LUAD using TCGA clinical survival genetic data [12]. Using the pathway information available from the GO-BP database to group genes into several pathways, we further conduct accurate survival predictions based on the selected main and interacting biomarkers.

**Methods**

We consider a study with $N$ independent subjects. For a subject $i$, suppose that there are $q$ environmental/clinical variables $\mathbf{e}_i = (e_{i1}, \ldots, e_{iq})'$, and $p$ genes $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ assigned to $G$ possibly overlapping pathways; that is, a given gene may belong to multiple
pathways. The pathway information accounts for the natural hierarchical structure of genes, and the overlapping pathways commonly exist in the gene expression data. Our aim is to determine the main features (genes and environment) and their interactions related to clinical survival outcomes, while taking into account the pathway information.

For a subject \( i \), assume the survival outcome \( t_i \) is related to the environmental/clinical variables \( e_i \), gene expression covariates \( x_i \), and their component-wise interactions \( w_i = (e_{i1}x_{i1}, \ldots, e_{i1}x_{ip}, e_{i2}x_{i1}, \ldots, e_{iq}x_{ip})' \) through the Cox regression model. In the Cox regression framework, the hazard function at time \( t \) for subject \( i \)'s survival given the covariates is modeled as

\[
\lambda(t|e_i, x_i, w_i) = \lambda_0(t) \exp(e_i'\alpha + x_i'\beta + w_i'\eta),
\]

where \( \lambda_0(t) \) is a non-negative deterministic baseline hazard function and \( (\alpha, \beta, \eta) \) are corresponding parameters. Usually the survival outcome is subject to censoring, and we use \( \delta_i \) to denote whether subject \( i \)'s survival time is observed or censored.

Incorporating the grouping (pathway) information into the modeling process may improve the interpretability and prediction accuracy of the model. When groups overlap with each other, special techniques are required to account for the overlapping grouping information. According to Jacob et al. [13], we decompose the original coefficient vector into the sum of group-specific potential effects, that is, \( \beta = \sum_{j=1}^{G} \gamma^j \) where \( \gamma^j = (\gamma^j_1, \ldots, \gamma^j_p)' \) is the latent coefficient vector for group \( j \). For \( j = 1, \ldots, G \) and \( k = 1, \ldots, p \), we set \( \gamma^j_k = 0 \) if gene \( k \) does not belong to group \( j \). Redefine the latent coefficient \( \gamma^j \) by removing the zero elements
therein, and form the latent coefficient vector \( \gamma \) by stacking the vectors \( \gamma^1, \ldots, \gamma^G \). Let \( u \) be the length of \( \gamma \). We can then rewrite \( \beta = S\gamma \), where \( S \) is a \( p \times u \) matrix whose elements are 1 or 0. A simple example for illustration is given in Appendix S.2.

On the basis of the coefficient decomposition, the original regression model can be transformed into a new model, that is, \( X_{N \times p} \beta_{p \times 1} = X_{N \times p} S_{p \times u} \gamma_{u \times 1} = \bar{X}_{N \times u} \gamma_{u \times 1} \), where \( \bar{X} = (x_1, \ldots, x_N)' \). Equivalently, this new model can be constructed by duplicating the columns of overlapping variables in the original design matrix. For the new transformed model, the hazard function for subject \( i \) in the Cox regression model is re-expressed as

\[
\lambda(t|\epsilon_i, \bar{x}_i, w_i) = \lambda_0(t)\exp(\epsilon_i'\alpha + \bar{x}_i'\gamma + w_i'\eta).
\]

The Method (OGS) for G-E interaction selection

We apply the OGS method to the environment and gene expression profile data with clinical survival trait to detect important main effects as well as interactions by incorporating prior pathway information. The steps of the OGS algorithm for G-E interaction selection are described as follows.

Step 1: We utilize the overlapping group Cox regression model to identify the candidate pathways based on the latent effect approach, which can be performed by the R package "grpregOverlap" [14]. We define \( \bar{M}_{\text{main}} \) as the selected set of pathways, and \( A = |\bar{M}_{\text{main}}| \) as the size of \( \bar{M}_{\text{main}} \).

Step 2: We utilize the sequence kernel association test (SKAT) to obtain the group-specific significance, where each group is formed by the interaction between the genes of each
candidate pathway selected in the first step and the environmental factors in Es, where Es is a set of environmental factors. Following Chen et al. [15], the SKAT statistic under the Cox regression model is defined as

\[ Q_k = r'R(k)W(k)W'(k)r, \quad k = 1, \ldots, A \]

Here, \( r \) is the vector of martingale residuals estimated from the null model by regressing survival outcomes on only the environmental covariates Es without considering the gene expression data; \( R(k) = [r(k)_{ij}]_{N \times l} \), where \( l \) is the number of G-E interaction pairs in the candidate pathway group \( k \), \( r(k)_{ij} \) is the \( j \)-th G-E interaction pair of \( i \)-th subject in the candidate pathway group \( k \), and \( W(k) \) is a diagonal weight matrix that contains the weights of the \( l \) interaction pairs in the candidate pathway group \( k \). Suitable weights can improve the testing power [16]. We utilize the penalized Cox partial likelihood approach with the Ridge penalty to estimate effect sizes for G-E interaction pairs in each candidate pathway group, and take the square root of the absolute estimated coefficients as our weights. Based on the null model by regressing survival outcomes on only the environmental covariates Es without gene covariates, let \( E \) is an \( N \times q \) design matrix for the \( q \) environmental covariates, and \( V = \text{diag}(c_1, \ldots, c_N) - PP' \), where \( P \) is an \( N \times v \) matrix with element \( p_{ij} \) the baseline hazard for individual \( i \) at ordered failure time \( t(j), \ j = 1, \ldots, v \), and \( c_i \) the cumulative hazard for individual \( i \) at observed time \( t_i \).

Let \( \Sigma(k) = W(k)R'(k)(V - VE(E'VE)^{-1}E'VE)R(k)W(k) \) be the covariance matrix of the vector \( W(k)R(k)r \) under the null hypothesis of all gene-environment interaction pairs in the
candidate pathway group $k$ having null effects. Under the null hypothesis, the SKAT statistic follows a mixture chi-square distribution:

$$Q_{(k)} \sim \sum_{j=1}^{l} \lambda_{(k)j} \chi_{1,j}^2,$$

where $\lambda_{(k)j}, j = 1, \ldots, l$ are the eigenvalues of $\Sigma_{(k)}$, and $\chi_{1,j}^2$'s are independent 1-df central chi-square random variables.

We use the Davies method [17] to approximate the tail probability of the mixture chi-square distribution, which can be calculated by the R package "CompQuadForm" [18]. Generally speaking, the Davies method is accurate [19]. The p-values $\{p_1, \ldots, p_A\}$ are used as our group screening measure; a smaller p-value corresponds to a higher group importance and therefore leads to a higher priority of selection.

Step 3: In the third step, we select significant G-E interactions based on the permutation procedure with the cutoff point determined by the soft-thresholding rule, where the permutation is applied to the covariate matrix consisting of both genes and environmental covariates. We randomly permute the original data $\{Y_i, e_i, x_i\}$ to form the permuted data $\{Y_i, e_{\pi(i)}, x_{\pi(i)}\}$ following the null model, where $Y_i = (t_i, \delta_i)$ is the survival outcome, and $\{\pi(1), \ldots, \pi(N)\}$ is a random permutation of the index. Then we apply again the SKAT test for each of the candidate pathway groups with the permuted data to obtain the group screening measures (p-values) $\{p_1^*, \ldots, p_A^*\}$. We adopt $C_{int} = \min\{p_1^*, \ldots, p_A^*\}$ as a cutoff point to select candidate pathway groups, i.e.

$$\hat{M}_{int} = \{b: p_b < C_{int}, b = 1, \ldots, A\},$$
is our selected set of candidate pathway groups.

Step 4: Finally, in the framework of joint modeling, based on environmental covariates, and selected genes and G-E interactions, a penalized regression with an appropriate penalty is used to establish the final survival prediction model. Therefore, we apply the penalized Cox's regression together with the Ridge or Lasso penalty to build the final prediction model based on all environmental variables, genes in $\tilde{M}_{main}$ and G-E interactions in $\tilde{M}_{int}$. Please note that when applying the Ridge penalty, all candidate biomarkers will be retained, and when applying the Lasso penalty, some candidate biomarkers may be removed by estimating their corresponding coefficients to be 0. The penalized Cox regression model with the Ridge or Lasso penalty can be obtained through the R package "glmnet" [20].

Results

Comparison with alternative methods in variable selection, estimation, and prediction

In the following simulations, we study the performances of the proposed OGS approach in variable selection, estimation and prediction, and compare them with the performance of the "Oracle", "Univariate Selection" and "Ordinary Lasso" methods. The "Oracle" method is based on the underlying true model, which is known in the simulations but unknown in real applications. The "Univariate Selection" method uses univariate regression to select environmental variables, genes, and G-E interactions one by one, with a controlled false discovery rate (<0.2), and then includes the selected variables in a multivariate Cox regression model to form the final prediction model. The "Ordinary Lasso" method is the penalized Cox
regression model with the Lasso penalty considering all environmental variables, genes, and G-E interactions in the model.

For performance comparison, we adopt the root mean squared error (RMSE) to measure estimation accuracy, defined as

\[
RMSE = \sqrt{\frac{1}{S} \sum_{j=1}^{S} (\theta_j - \hat{\theta}_j)^2}
\]

where \( S \) is the size of the full model including all main and interaction covariates and \( \theta' = (\alpha', \beta', \eta') \).

To evaluate the estimation performance, we report RMSE.M, the mean of the root mean square errors of 200 simulations. To evaluate the performance of the selection accuracy, we consider various criteria: T.model is the proportion of the selected models over 200 simulations containing all the underlying effective variables, including both the main and interaction terms; Tint.model is the proportion of the selected models over 200 simulations containing all the underlying effective G-E interaction terms; Sen. is the sensitivity, defined as the proportion of the underlying effective variables being selected; Spe. is the specificity, defined as the proportion of the underlying ineffective variables not being selected. We also report the average size of the selected models, S.model, in 200 simulations. To evaluate the performance of survival prediction, we consider two measures of prediction accuracy: the deviance and the c-index proposed by Harrell et al. [21]; smaller deviance or larger c-index corresponds to better prediction accuracy. The mean values of deviance and c-index over 200 simulations are reported.
Let $\hat{\theta}' = (\hat{\alpha}', \hat{\beta}', \hat{\eta}')$ an estimator of the (penalized) Cox regression parameter in a prediction model obtained from the training dataset. Let $(t_i, \delta_i, e_i', x_i', w_i')$ be the survival and covariate data of subject $i$ in the test data. Define $(e_i', x_i', w_i')\hat{\theta}$ as the prognosis index (PI) value for subject $i$ in the test data. The Cox test is defined as the p-value of PI when PI is used as the covariate in the univariate Cox model for survival outcomes in the test data. Similarly, the LR-test is the p-value of the log-rank test for the null hypothesis of equal survival between the “good” and “poor” prognostic groups in the test data, where the “good” and “poor” prognostic groups are classified according to whether the PI value is higher or lower than the median PI value in the test data. Smaller Cox-test and LR-test values correspond to better predictive power.

In simulations we consider survival data with a cohort size 300 in the training set, where each subject’s survival time follows the Cox proportional hazards model

$$\lambda_0(t|e, x, w) = 10\exp(e'\alpha + x'\beta + w'\eta),$$

with the covariates $e$ and $x$ jointly following a multivariate standard normal distribution with correlation $\text{corr}(e_j, e_k) = 0.3|j-k|$, $\text{corr}(x_j, x_k) = 0.5|j-k|$, and $\text{corr}(e_j, x_k) = 0$ for all $j, k$. The censoring time distribution follows a uniform distribution. We then generate survival data, independent of the training data, with a cohort size 100 as the test data to assess the prediction accuracy for different methods.

In this simulation study, we consider 5 environmental variables and assume that the first 4 are related to the survival outcome, and the corresponding effects are $1.5, 2.25, 3, -1.5$. On
the other hand, the gene covariates considered contain 25 groups that have different group sizes (the numbers of genes) and may share with each other some of the genes. The group sizes and the overlapping structure (i.e. the number of the shared genes between two overlapping groups) are shown in Table 1, where the overlapping groups are shown side by side. For example, group 1 contains 3 genes, as group 2 does, but the two groups contain only 5 unique genes, and 1 gene is shared between the two groups. As a result, there are a total of 500 genes and 632 group-specific latent effects (see Section 2) in this example. Fig. 2 displays the gene network structure. Groups 1, 7, 13, and 19 are set to be effective, and genes in each of them have constant latent effects of 3, 3, 2, and -2, respectively. In addition, effective interactions (E1*G22, E1*G24, E2*G26) with the corresponding effects (4.5,4.5,6) and (E2*G78, E3*G83, E3*G88) with the corresponding effects (−3, −4.5, −6) are in group 7 and group 13, respectively. The number of effective environment, gene, or G-E interaction factors is 91 among a total of 3,005 such factors. We examine the performances of different methods under a censoring rate of 30%, 50%, or 70%. We also conduct further simulations to demonstrate the performances of the new proposal, whose details and results can be seen in Appendix S.3.

| Pathway | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Gene Size | 3 | 3 | 6 | 6 | 9 | 9 | 15 | 15 | 15 | 24 | 24 | 24 | 36 | 36 | 36 | 45 | 45 | 45 | 60 | 60 | 60 | 60 | 38 |
| Overlapping | 1 | 1 | 0 | 2 | 2 | 0 | 3 | 3 | 0 | 5 | 5 | 0 | 8 | 8 | 0 | 12 | 12 | 0 | 15 | 15 | 0 | 20 | 20 | 0 |
Summary of simulation results

From the simulation results shown in Table 2 and Table S.3 in Appendix S.3 where the gene network structure is complex, we see that the OGS method using the Lasso or Ridge penalty performs substantially better than the "Univariate Selection" and the "Ordinary Lasso" methods in variable selection, effect estimation, and survival prediction. On the other hand, simulation results shown in Table S.4 of Appendix S.3 where a simpler gene structure is considered, the performance of OGS with Lasso or Ridge penalty is worse than that of the "Ordinary Lasso" method when the censoring rate is 30% or 50%; while when the censoring rate is higher (70%), the OGS with the Lasso or Ridge penalty performs better than the "Ordinary Lasso".

Furthermore, further simulation studies with a small cohort size are conducted under the scenario where all simulated settings are the same as those in the previous simulation study
except for a cohort size defined as 150/50 (training/testing). We still obtain similar numeric results patterns; these corresponding results are shown in Table S.5~S.7 in Appendix S.3.

**Table 2** The average of the performance measures out of 200 simulation replications for different approaches

|                | Oracle | Uni. Sel. | Ordinary Lasso | OGS Ridge | OGS Lasso |
|----------------|--------|-----------|----------------|-----------|-----------|
| **Censoring rate = 30%** |        |           |                |           |           |
| RMSE           | 0.4039 | 0.4196    | 0.4128         | 0.4106    | 0.3830    |
| T.model        | 1.0000 | 0.0000    | 0.0000         | 0.3650    | 0.3050    |
| Tint.model     | 1.0000 | 0.0000    | 0.2250         | 0.3900    | 0.3900    |
| Sen.           | 1.0000 | 0.2507    | 0.6956         | 0.9723    | 0.9114    |
| Spe.           | 1.0000 | 0.9978    | 0.9829         | 0.9488    | 0.9920    |
| C.model        | 91.0000| 29.1600   | 113.1650       | 237.8050  | 106.1550  |
| Deviance       | -120.1966 | -24.9825 | -93.6840       | -106.4978 | -191.8980 |
| C-index        | 0.8661 | 0.7092    | 0.8430         | 0.8835    | 0.9126    |
| **Censoring rate = 50%** |        |           |                |           |           |
| RMSE           | 0.3946 | 0.4179    | 0.4159         | 0.4117    | 0.4021    |
| T.model        | 1.0000 | 0.0000    | 0.0000         | 0.3250    | 0.0100    |
| Tint.model     | 1.0000 | 0.0050    | 0.0450         | 0.3800    | 0.2850    |
| Sen.           | 1.0000 | 0.2577    | 0.5524         | 0.9671    | 0.8172    |
| Spe.           | 1.0000 | 0.9970    | 0.9838         | 0.9403    | 0.9933    |
| C.model        | 91.0000| 32.1550   | 97.4050        | 262.0400  | 94.0050   |
| Deviance       | -91.7132 | -17.1172 | -55.4557       | -79.1409  | -108.8120 |
| C-index        | 0.8803 | 0.7278    | 0.8194         | 0.8883    | 0.8925    |
| **Censoring rate = 70%** |        |           |                |           |           |
| RMSE           | 0.3898 | 0.4198    | 0.4188         | 0.4138    | 0.4114    |
| T.model        | 1.0000 | 0.0000    | 0.0000         | 0.1800    | 0.0000    |
| Tint.model     | 1.0000 | 0.0000    | 0.0000         | 0.2800    | 0.0500    |
| Sen.           | 1.0000 | 0.1666    | 0.3847         | 0.8668    | 0.5627    |
| Spe.           | 1.0000 | 0.9973    | 0.9855         | 0.9503    | 0.9955    |
| C.model        | 91.0000| 23.1000   | 77.1300        | 223.6400  | 64.3250   |
| Deviance       | 22.7686 | -1.6320   | -24.1135       | -46.6483  | -45.5560  |
| C-index        | 0.8517 | 0.7008    | 0.7826         | 0.8709    | 0.8459    |
**Real data application: TCGA HNSCC data**

The TCGA HNSCC RNA-Seq expression data, which were collected using the IlluminaHiseq RNAseq V2 platform, together with the phenotype data containing the survival time and censoring status data for 484 subjects, can be downloaded from the R package “GEInter” [22]. The censoring rate of the survival time is about 58%, and gene expression measurements for a total of 18,409 genes are available in this data. As the number of cancer-related genes is not expected to be large, we conduct prescreening using marginal Cox models, which can also improve stability for feature selection. The top 2,000 genes with the smallest p-values in the marginal (univariate) Cox models are selected for downstream analysis. The five E factors analyzed including age, gender, AJCC pathologic stage nodes, AJCC pathologic stage tumor, and ICD O3 site. Summary information for these clinical variables is reported in the Table 3.

| Variable             | Coding                        | Missing status | Continuous(EC) /Discrete(ED) |
|----------------------|-------------------------------|----------------|-------------------------------|
| age                  | No                            | EC             |                               |
| gender               | female=0, male=1              | No             | ED                            |
| AJCC pathologic nodes| n0=0, n1=1, (n2, n2a, n2b, n2c)=2, n3=3, nx=4 | No             | ED                            |
| AJCC pathologic tumor| t0=0, t1=1, t2=2, t3=3, (t4, t4a, t4b)=4, tx=5 | No             | ED                            |
| ICD O3 site          | (C00.9, C01.9, C02.1, C02.9)=0, (C03.0, C03.1, C03.9, C04.0, C04.9)=1, | No             | ED                            |
The PTReg method [5] was developed to conduct robust joint analysis using penalized trimmed regression with the MCP penalty under the AFT model for the right-censored survival outcome. We are interested in comparing the PTReg approach with our proposed OGS approach in the real data application. The whole 12,005 main and G-E interaction predictors are considered for the "Univariate Selection", "Ordinary Lasso", and "PTReg" methods. For the OGS method, among the 2,000 preselected genes, prior pathway information for 1,637 genes, which are mapped into 4,651 pathways based on the GO biological process database, is utilized. The 363 genes that are not mapped into any pathways in the GO biological process database are either discarded or put together as a group for the latent effect analysis in the OGS method, leading to a total of 9,827 or 12,005 main and G-E interaction effects considered.

We take ten random splits of the whole data into 387:97 training/test sets to evaluate the performances of all the methods considered in the TCGA HNSCC data application. Table 4 reports the median of the survival prediction results over the ten folds when the 363 ungrouped genes are discarded from analysis. We see that the performance of the OGS method with Ridge or Lasso penalty is better than the "Univariate Selection", "Ordinary Lasso", and "PTReg" methods. The OGS approach putting the 363 ungrouped genes together as an additional group
results in the same prediction model as the one discarding the ungrouped genes. Also, the OGS analysis results based on the pathway information obtained from other annotated gene set databases, including GO cellular component (GO-CC), GO molecular function (GO-mf), KEEG, and Biocarter, are compared with the other methods for survival prediction in the TCGA HNSCC data, as shown in Table S.8. These additional results based on pathway information from alternative gene set databases still reveal that the OGS approach performs better than the other methods.

Table 4

|                | Uni. Sel. | Ordinary_Lasso | OGS_Ridge | OGS_Lasso | PTReg |
|----------------|----------|----------------|-----------|-----------|-------|
| Cox-test       | 0.2181   | 0.0844         | 0.0056    | 0.0074    | 0.0261|
| LR-test        | 0.1411   | 0.0999         | 0.0200    | 0.0301    | 0.1320|
| Deviance       | 1501.3094| 14.0777        | 9.0998    | 9.5429    | 57.4644|
| C-index        | 0.5570   | 0.5658         | 0.6452    | 0.6354    | 0.5665|

Based on one random split of the data, Fig. 3 displays the Kaplan-Meier survival curves of the “good" and “poor" prognosis groups in the test data. It can be seen that the OGS method separates the two groups better than other methods. When applying the OGS with the Lasso penalty to the entire data based on the GO biological process database, we identify several significant main and G-E interaction effects, and obtain the corresponding parameter estimates, as shown in Table S.9. We note that the clinical variable "AJCC pathological stage tumor" interacts with several genes, and most of
these genes, such as LHX1 [23] and SZT2 [24-25], have been shown to be related to HNSCC.

Fig. 3

Kaplan-Meier curves for the 97 subjects in the TCGA HNSCC testing data. Good and poor groups are identified by the median of the PI scores in the test dataset

Real data application: TCGA ESCA data

The TCGA ESCA RNA-Seq expression data, together with the phenotype data containing the survival time and censoring status data, can be downloaded from the R package ’TCGAbiolinks’ [26], or ’UCSCXenaTools’ [27]. After excluding patients with missing survival time data, our analysis is focused on the subset of the TCGA ESCA data with
368 patients and 20,501 gene expression variables. The censoring rate of the survival time in
the data is about 58%. The TCGA ESCA clinical information data can be obtained from
the ‘FireBrowse’ database [28].

Since the number of cancer-related genes is expected to be limited, when applying the
proposed OGS method, the top 2,000 genes with the smallest p-values based on the marginal
(univariate) Cox models are selected for downstream analysis. The seven clinical variables
whose E effects are analyzed include age, gender, esophageal tumor central location, peson
neoplasm cancer status, race, BMI, and AJCC pathologic stage, and their summary
information is reported in the Table 5. Some of the clinical variables contain missing values,
and we use the sparse boosting method [29] in the R package "GEInter" to perform multiple
imputation for the missing values in the clinical variables. Based on the GO biological process
database, 1,440 genes among the top 2,000 genes are mapped into 4,290 pathways and such
prior pathway information is utilized in the OGS method. Excluding the genes without being
mapped into any pathway, there are a total of 11,527 main and G-E interaction covariates in
the proposed OGS method. On the other hand, a total of 16,007 main and G-E interaction
predictors are considered in the "Univariate Selection", "Ordinary Lasso", and "PTReg"
methods.

| Variable                  | Coding          | Missing status | Continuous(EC) /Discrete(ED) |
|---------------------------|-----------------|----------------|------------------------------|
| esophageal tumor central  | proximal=1,     | Yes            | ED                           |
|                           | mid=2,          |                |                              |
| location | distal=3 |
|----------|----------|
| person   | tumor free=1, with tumor=2, Yes ED |
| neoplasm | cancer status |
| race     | white=1, asian=2, black or African american=3 Yes ED |
| BMI      | weight/height^2 Yes EC |
| AJCC     | pathologic stage |
|          | (stage i, stage ia, stage ib) =1 |
|          | (stage ii, stage iia, stage iib)=2 |
|          | (stage iii, stage iia, stage iib, stage iiic)=3 |
|          | (stage iv, stage iva)=4 |
| age      | days_to_birth No EC |
| gender   | female=0, male=1 No ED |

We take ten random splits of the whole TCGA ESCA data into 294:74 training/test sets to evaluate the performances of all methods for survival prediction in the TCGA ESCA data. Table 6 reports the median of the survival prediction results among the ten folds. We see that the performance of the OGS method with the Ridge or Lasso penalty is better than the "Univariate selection", "Ordinary Lasso", and "PTReg" methods. In addition to the OGS analysis discarding the 560 genes without mapped pathways in the GO biological process database, we also perform the OGS analysis putting the unmapped genes together as an additional group, and the two different implements of the OGS method result in the same prediction model. Also, different annotated gene sets databases, including GO-CC, GO-MF, KEEG, and Biocarter are also used in the OGS approach to catch pathway information. As
shown in Table S.10. the OGS method still outperforms than the other methods using such alternative pathway information.

|         | Uni. Sel. | Ordinary_Lasso | TS-OGS Ridge | TS-OGS Lasso | PTReg |
|---------|-----------|----------------|--------------|--------------|-------|
| Cox-test| 0.0190    | 2.0732e-08     | 1.4088e-08   | 2.5767e-09   | 0.0023|
| LR-test | 0.1561    | 2.1075e-05     | 6.8139e-08   | 1.2715e-07   | 0.0070|
| Deviance| 256.2041  | -14.0819       | -19.9024     | -27.2311     | 63.8834|
| C-index | 0.5773    | 0.7935         | 0.8551       | 0.8564       | 0.7194|

Based on one random split of the data, Fig. 4 displays the Kaplan-Meier survival curves for the “good" and “poor" prognosis groups in the test data. It is seen that the two survival curves are better separated by the OGS approach than other methods. When applying the OGS with the Lasso penalty for whole data based on the GO biological process database, we identify and estimate several significant main and G-E interaction effects, which are shown in Table S.11. We find that the two specific genes TSPYL2 (Testis Specific Protein Y-Linked 2) and TSPYL4 (Testis Specific Protein Y-Linked 4) interact with the clinical variables “BMI”, “AJCC pathologic stage", and “age".
Kaplan-Meier curves for the 74 subjects in the TCGA ESCA testing data. Good and poor groups are identified by the median of the PI scores in the test dataset.

In addition to the previous analyses based on TCGA data, the analysis results for TCGA LUAD data are provided in Appendix S.4.3. We find that the four specific genes, DKK1, VAX1, EPGN, and EREG, interact significantly with most of the clinical variables mentioned above. In fact, the proposed OGS approach performs consistently well across these different cancer datasets.
Conclusion

In this article, we propose a two-stage overlapping group screening procedure to identify important main and gene-environment (G-E) interaction effects efficiently for survival prediction. In the first stage, the new proposal utilizes the latent effect approach to identify candidate gene pathways for survival prediction, adjusting for the E and G-E interaction factors. Different gene pathways are allowed to overlap with each other, i.e., to share common genes. In the second stage, we utilize the SKAT approach [15], which is a popular group testing approach, to obtain the group-level p-value of each candidate gene pathway as well as the associated G-E factors, adjusting for the E factors. A pathway as well as the associated G-E factors is then selected when their group-level p-value is smaller than the one under covariate (both G and E factors) permutation. The final survival prediction model is constructed by a Cox model based on the E factors, the selected gene pathways as well as the associated G-E factors, subject to the Ridge or Lasso penalty. Simulation and real data studies demonstrate that, compared with the analysis that ignores pathway information, the new proposal can significantly improve the accuracy of gene and gene-environment interaction selection, as well as the resulting survival predictions.

Discussion

The OGS method is flexible. Although we focus on survival prediction based on the Cox proportional hazards model, the same idea can straightforwardly apply to other outcome models, such as the proportional odds survival time model, the logistic regression model for
binary outcomes, and the multinomial logistic regression model for multi-class outcomes. For example, the SKAT statistic involved in the OGS method can be modified simply by using the residuals from the alternative model under consideration.

The OGS method employs the latent effect approach to extract gene network structure information in terms of gene pathways. This requires a pre-designated gene group (pathway) structure and is limited to genes that can be assigned to at least one group (pathway). It is interesting to study how to relax these restrictions to improve the performance of feature selection and survival prediction in the presence some covariate network structure.

Abbreviations

AFT: Accelerated failure time; AJCC: American joint committee on cancer; BMI: Body mass index; ESCA: Esophageal carcinoma; GEInter: Gene–environment interaction; G-E: Gene–environment; GO-BP: Gene Ontology biological process; GO-CC: Gene Ontology cellular component; GO-MF: Gene Ontology molecular function; HNSCC: Head and neck squamous cell carcinoma; Lasso: Least absolute shrinkage and selection operator; LR-test: Log-rank test; LUAD: Lung adenocarcinoma; MCP: Minimax concave penalty; OGS: Overlapping grouped screening; PTReg: Penalized trimmed regression; PI: Predictor index; RMSE: Root mean squared error; SKAT: Sequence kernel association test; SNPs: Single-nucleotide polymorphisms; TCGA: The Cancer Genome Atlas.

Declarations

Author’s contributions
Conceived and designed the experiments: JH. Analyzed the data: JH, KH. Wrote the first draft of the manuscript: JH. Contributed to the writing of the manuscript: JH, YH. Agree with manuscript results and conclusions: JH, KH, YH. Jointly developed the structure and arguments for the paper: JH, KH, YH. Made critical revisions and approved final version: JH, YH. All authors reviewed and approved of the final manuscript.

Acknowledgements

None

Funding

This research is supported by the grant MOST 110-2118-M-035-001- MY2 from the Ministry of Science and Technology of Republic of China (Taiwan).

Availability of data and materials

The TCGA ESCA, and LUAD genomic data with survival traits and pathway information database analyzed during this study are all available at figshare website https://figshare.com/articles/dataset/TCGA_cancer_genomic_data/16816654. The TCGA HNSCC genomic data can be downloaded from the R package “GEInter” [22]. R codes for the simulation studies and real data are available at figshare website https://figshare.com/articles/software/R_codes_for_simulation_and_real_data/16816303.

Ethics approval and consent to participate

Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Reference

1. McAllister K, Mechanic LE, Amos C, et al. Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am J Epidemiol*. 2017;186(7):753-761. doi:10.1093/aje/kwx227

2. Thomas D. Gene--environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010;11(4):259-272. doi:10.1038/nrg2764

3. Batchelor TT, Betensky RA, Esposito JM, et al. Age-dependent prognostic effects of genetic alterations in glioblastoma. *Clin Cancer Res*. 2004;10(1 Pt 1):228-233. doi:10.1158/1078-0432.ccr-0841-3

4. Zhou F, Ren J, Lu X, Ma S, Wu C. Gene-Environment Interaction: A Variable Selection Perspective. *Methods Mol Biol*. 2021;2212:191-223. doi:10.1007/978-1-0716-0947-7_13

5. Xu Y, Wu M, Zhang Q, Ma S. Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics*. 2019;111(5):1115-1123. doi:10.1016/j.ygeno.2018.07.006

6. Xu Y, Wu M, Ma S, Ahmed SE. Robust gene-environment interaction analysis using
penalized trimmed regression. *J Stat Comput Simul.* 2018;88(18):3502-3528. doi:10.1080/00949655.2018.1523411

7. Wu M, Zhang Q, Ma S. Structured gene-environment interaction analysis. *Biometrics.* 2020;76(1):23-35. doi:10.1111/biom.13139

8. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. B.* 2008;70(5):849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

9. Wang JH, Chen YH. Network-adjusted Kendall's Tau Measure for Feature Screening with Application to High-dimensional Survival Genomic Data [published online ahead of print, 2021 Jan 29]. *Bioinformatics.* 2021;btab064. doi:10.1093/bioinformatics/btab064

10. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

11. Wang, JH., Chen, YH. Overlapping group screening for detection of gene-gene interactions: application to gene expression profiles with survival trait. *BMC Bioinformatics* **19**, 335 (2018). https://doi.org/10.1186/s12859-018-2372-2

12. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008;455(7216):1061-1068. doi:10.1038/nature07385

13. Jacob L, Obozinski G, Vert JP. Group lasso with overlap and graph lasso. In: Proceedings
of the 26th annual international conference on machine learning. Montreal: ACM; 2009. p. 433–40.

14. Zeng Y, Breheny P. Overlapping Group Logistic Regression with Applications to Genetic Pathway Selection. *Cancer Inform.* 2016;15:179-187. doi:10.4137/CIN.S40043

15. Chen H, Lumley T, Brody J, et al. Sequence kernel association test for survival traits. *Genet Epidemiol.* 2014;38(3):191-197. doi:10.1002/gepi.21791

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

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

18. Davies RB, Algorithm AS. 155: The distribution of a linear combination of $x^2$ random variables. *J R Stat Soc Ser C Appl Stat.* 1980;29(3):323–33.

19. Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput Stat Data Anal.* 2010;54(4):858–62.

20. Wu B, Guan W, Pankow JS. On Efficient and Accurate Calculation of Significance P-Values for Sequence Kernel Association Testing of Variant Set. *Ann Hum Genet.* 2016;80(2):123-135. doi:10.1111/ahg.12144
21. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw.* 2011;39(5):1-13. doi:10.18637/jss.v039.i05

22. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15(4):361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

23. Wu M, Qin X, Ma S. GEInter: an R package for robust gene-environment interaction analysis. *Bioinformatics.* 2021;37(20):3691-3692. doi:10.1093/bioinformatics/btab318

24. Zhang, B., Wang, H., Guo, Z. *et al.* A panel of Transcription factors identified by data mining can predict the prognosis of head and neck squamous cell carcinoma. *Cancer Cell Int* 19, 297 (2019). [https://doi.org/10.1186/s12935-019-1024-6](https://doi.org/10.1186/s12935-019-1024-6)

25. Tan FH, Bai Y, Saintigny P, Darido C. mTOR Signalling in Head and Neck Cancer: Heads Up. *Cells.* 2019;8(4):333. doi:10.3390/cells8040333

26. Sakai A, Ando M, Fukusumi T, et al. Aberrant expression of CPSF1 promotes head and neck squamous cell carcinoma via regulating alternative splicing. *PLoS One.* 2020;15(5):e0233380. doi:10.1371/journal.pone.0233380

27. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 2016;44(8):e71. doi:10.1093/nar/gkv1507
28. Wang S, Liu X. The UCSCXenaTools R package: a toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq. *J. Open Source Softw.* 2019;4(40):1627. [https://doi.org/10.21105/joss.01627](https://doi.org/10.21105/joss.01627)

29. Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, Perner S. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database (Oxford)*. 2017;2017:baw160. Published 2017 Jan 6. doi:10.1093/database/baw160

30. Wu M, Ma S. Robust semiparametric gene-environment interaction analysis using sparse boosting. *Stat Med.* 2019;38(23):4625-4641. doi:10.1002/sim.8322
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BMCOGSAAppendix.docx