frontiers in Genetics

ORIGINAL RESEARCH
published: 02 August 2021
doi: 10.3389/fgene.2021.701405

Cox-sMBPLS: An Algorithm for Disease Survival Prediction and Multi-Omics Module Discovery Incorporating *Cis*-Regulatory Quantitative Effects

*Nasim Vahabi[1], Caitrin W. McDonough[2], Ankit A. Desai[3], Larisa H. Cavallari[2], Julio D. Duarte[2] and George Michailidis[1]**

[1] Informatics Institute, University of Florida, Gainesville, FL, United States, [2] Department of Pharmacotherapy and Translational Research, Center for Pharmacogenomics and Precision Medicine, University of Florida, Gainesville, FL, United States, [3] Department of Medicine, Indiana University, Indianapolis, IN, United States

Background: The development of high-throughput techniques has enabled profiling a large number of biomolecules across a number of molecular compartments. The challenge then becomes to integrate such multimodal Omics data to gain insights into biological processes and disease onset and progression mechanisms. Further, given the high dimensionality of such data, incorporating prior biological information on interactions between molecular compartments when developing statistical models for data integration is beneficial, especially in settings involving a small number of samples.

Results: We develop a supervised model for time to event data (e.g., death, biochemical recurrence) that simultaneously accounts for redundant information within Omics profiles and leverages prior biological associations between them through a multi-block PLS framework. The interactions between data from different molecular compartments (e.g., epigenome, transcriptome, methylome, etc.) were captured by using *cis*-regulatory quantitative effects in the proposed model. The model, coined Cox-sMBPLS, exhibits superior prediction performance and improved feature selection based on both simulation studies and analysis of data from heart failure patients.

Conclusion: The proposed supervised Cox-sMBPLS model can effectively incorporate prior biological information in the survival prediction system, leading to improved prediction performance and feature selection. It also enables the identification of multi-Omics modules of biomolecules that impact the patients' survival probability and also provides insights into potential relevant risk factors that merit further investigation.

Keywords: multi-omics, supervised Integration, *cis*-regulatory quantitative, multi-block PLS, survival analysis

OPEN ACCESS

Edited by:
Gong Zhang,
Jinan University, China

Reviewed by:
Shihua Zhang,
Academy of Mathematics and Systems Science (CAS), China
Huaizhen Qin,
University of Florida, United States

****Correspondence:***
George Michailidis
gmichail@ufl.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *27 April 2021*
Accepted: *07 July 2021*
Published: *02 August 2021*

Citation:
Vahabi N, McDonough CW, Desai AA, Cavallari LH, Duarte JD and Michailidis G (2021) Cox-sMBPLS: An Algorithm for Disease Survival Prediction and Multi-Omics Module Discovery Incorporating Cis-Regulatory Quantitative Effects. Front. Genet. 12:701405. doi: 10.3389/fgene.2021.701405

INTRODUCTION

A key aim in integrating multi-Omics data is to identify combinations of molecular biomarkers that are either predictive of disease onset and outcomes or lead to insights into biological processes and disease mechanisms. To achieve this data integration, it is important to leverage information on interactions/mediations of the different molecular compartments profiled and measured by the

various Omics technologies. For example, DNA methylation is known to influence the phenotypic outcome of genetic variation and offers highly complementary information on transcriptional silencing and gene imprinting (Kass et al., 1997). To characterize associations between epigenomics, genomics, and transcriptomics, identification of *cis*-regulatory quantitative effects of SNPs on DNA methylation (meQTL) and mRNA expression (eQTL), and the effect of DNA methylation on mRNA expression (eQTM) has proved particularly informative (Jones, 1999).

In the past decade, a large body of literature was developed to introduce methods relating Omics profiles and disease outcomes, such as recurrence in cancer patients, death, etc. (Park et al., 2002; Tan et al., 2006; Zhang and Zhang, 2020). The most widely used method to model the time to such events is the Cox proportional hazard (Cox-PH) model (Cox, 1972), for which a number of adaptations have been proposed in the literature to make it suitable for use in high-dimensional settings induced by Omics data. Some adaptations leverage various variable selection methods –stepwise approaches, or regularization methods (see Bühlmann et al., 2013 and references therein)-, while others focus on reducing the dimensionality of the predictors by using principal components analysis (PCA) or partial least squares (PLS) (Wold et al., 1983). Partial least squares regression for non-numerical outcome variables was introduced in Garthwaite (1994) and Bastien and Tenenhaus (2001) (PLS-Generalized Linear Regression), and for survival data in (Bastien et al., 2015).

Ridge regression (Hoerl et al., 1975), as the first generation of L_P-regularization methods, utilizes the $L_2 - norm$ of the regression coefficients to improve prediction performance. However, ridge regression only shrinks the coefficients toward zero. Instead the Lasso method (Tibshirani, 1996) aims to simultaneously shrink and select a subset of variables through an $L_1 - norm$ constraint on the regression coefficients. An important limitation of the lasso method, especially in the case of Omics data, is that lasso tends to select only one variable among a group of correlated variables. For instance, in the multi-Omics framework, there are many features which are interacting as a network (or module) and sharing the same biological pathway. Therefore, the lasso method can poorly indicate this grouping information in the multi-Omics setting. Theoretical and practical explanations of this limitation are given in Efron et al. (2004), and Zou and Hastie (2005). To address these limitations, the elastic-net (Zou and Hastie, 2005) was introduced by imposing a convex combination of the lasso and ridge (L_1, L_2) penalties on the regression coefficients including Cox model. A recent benchmark analysis (Jardillier et al., 2020) of lasso-like penalties (including ridge, lasso, adaptive lasso, and elastic-net) of the Cox model showed a better prediction performance of elastic-net Cox compare to lasso-Cox and adaptive lasso-Cox models.

PLS regression (Wold et al., 1983) has also been used as a dimension reduction method in high-dimensional settings. This method is extended by Garthwaite (1994), and Bastien and Tenenhaus (2001), for generalized regression models (PLS-GLR) and the Cox-PH model as a special case (without considering censoring information). Further developments of PLS-GLR are introduced by Bastien et al. (2005). Chun and Keleş (2010), showed that a large number of features in the high-dimensional framework could greatly affect the prediction performance in PLS regressions. They proposed the sparse PLS (sPLS) by incorporating a variable selection constraint directly on the PLS direction vectors (weights). Lee et al. (2013), proposed a new formulation of the sPLS algorithm for survival data. Thereafter, Bastien et al. (2015), proposed a new algorithm called sparse PLS deviance residual (sPLSDR) by use of the normalized martingale residuals as the response variable in the sPLS algorithm.

Random forest, RF (Breiman, 2001), is another powerful prediction system that can consider more complex dependencies between the features. Random survival forest, RSF (Ishwaran et al., 2008), is an extension of random forest RF to analyze survival data (in the presence of right censoring) by introducing a new splitting rule and missing data imputation algorithm. RSF has shown reliable predictions in single-Omics settings (refer to Yosefian et al., 2015 and the references therein). Block Forest (Hornung and Wright, 2019) is another recent extension of RF which considers multiple tuning parameters, where each tuning parameter is associated with one of the data blocks. However, these approaches do not distinguish between variables obtained from different molecular compartments and also ignore any biological constraints –e.g., many variables belong to the same functional pathway, or act as regulators of other variables.

A similar issue arises in complex chemical systems, where variables can be naturally grouped into blocks. Multi-Block PLS, MBPLS (Wangen and Kowalski, 1989) was developed to study association between a numerical outcome variable and blocks of *a priori* defined predictors. The algorithm estimates the model parameters for each block and combines them using the relative importance of each block in predicting the outcome variable. MBPLS has been mostly employed in chemistry, however, multi-Omics data provide a novel opportunity to further extend and apply this algorithm in bioinformatics and genomics (Li et al., 2012).

In this paper, we propose a new integrative survival prediction model named supervised Cox sparse Multi-Block Partial Least Squares (Cox-sMBPLS) by simultaneously controlling the redundancy between Omics profiles from different molecular compartments, focusing on epigenomics, genomics and transcriptomics and incorporating *cis*-regulatory quantitative effects (eQTL, eQTM, meQTL) to integrate additional biological information to the training of the model. Note that the model and integrative strategy are general and can be easily adapted to other molecular compartments with appropriate modifications.

To handle censoring in the survival outcome data, we employ a reweighting technique described in the section "Materials and Methods." The high dimensionality of the Omics data under consideration is dealt with the use of regularization. The key objective of Cox-sMBPLS is to determine multi-Omics modules [i.e., genes, single-nucleotide polymorphisms (SNPs), and cytosine-phosphate-guanine (CpGs)

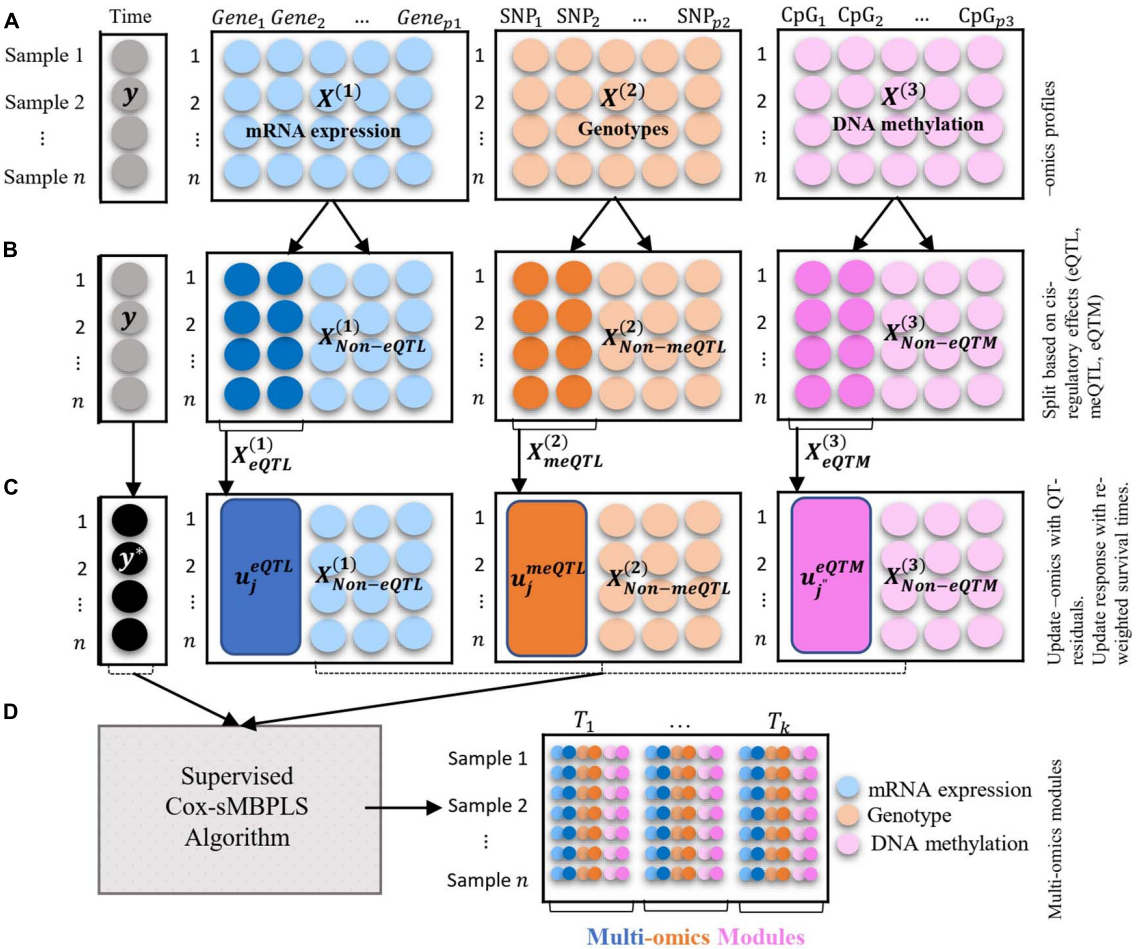


FIGURE 1 | Illustration of the multi-blocks data structure and the supervised Cox-sMBPLS algorithm. **(A)** three blocks of multi-Omics profiles: mRNA expression, genotypes, and DNA methylation. The response variable (**y**) is a vector of survival times of size $n = 1$. **(B)** splitting the Omics blocks based on the *cis*-regulatory quantitative effects (eQTL, eQTM, meQTL). **(C)** updating Omics blocks with QT-residuals, and **y** with reweighted times. **(D)** Multi-Omics modules which are combinations of all three Omics profiles.

sites], that are most associated with disease progression and patient's survival.

MATERIALS AND METHODS

An overview of the multi-block data structure used as input, together with how associations between molecular compartments are captured through *cis*-regulatory quantitative effects (eQTL, eQTM, meQTL) to extract low dimensional Omics modules that are predictive of survival times is depicted in **Figure 1**. As previously mentioned, to handle censoring information on the survival times, we employ an inverse censoring probability weighting scheme to adjust the response variable (observed survival time). The prediction and feature-selection performance of the model is evaluated using various metrics (see parameter tuning and model performance evaluation sub-section).

The rest of the "Materials and Methods" section is as follows: in section "Reweighted Survival Time" we introduce the reweighting of survival time (response variable) using an inverse censoring probability weighting scheme. Doing so we are considering the censoring information in the response variable as well. The reweighted survival time will be used as the response variable in the proposed Cox-sMBPLS algorithm. In section "Partial Least Squares Regression" we briefly introduce the partial least squares regression. In section "Integrating Cis-regulatory Quantitative Effects" we introduce the full process of integrating *cis*-regulatory quantitative effects (eQTL, eQTM, meQTL) and updating Omics-blocks $[X^{(b)},\ b = 1, 2, 3]$ by regressing out the shared information between QT-pairs (such as, gene-SNP in eQTL) and replacing that with the QT-residuals. In section "Supervised Cox-sMBPLS Algorithm" we provide our proposed supervised Cox-sMBPLS (supervised Cox sparse Multi-Block Partial Least Squares) which uses the reweighted survival time (section "Reweighted Survival Time") and updated Omics-blocks (section "Integrating Cis-regulatory Quantitative Effects") as outcome and covariates, respectively. In this section, we first introduce the objective function of

supervised Cox-sMBPLS and its solution in the case of univariate response (such as survival time) followed by the detailed algorithm implementation (**Algorithm 1**). In section "Parameter Tuning and Model Performance Evaluation" we explain the parameter tuning procedure and performance measures which are used to evaluate the feature-selection performance and probability of selecting the correct (important) features for the proposed model. Data sources are fully described in section "Data Source."

Reweighted Survival Time

Let $\tilde{y}_i$ and C_i indicate the independent true survival time and censoring time for the i^{th} subject ($i = 1, \ldots, n$), respectively. The observed data consist of pairs $\left\{\left(y_i, \delta_i\right) | i = 1, \ldots, n\right\}$ where $y_i = min\left(\tilde{y}_i, C_i\right)$ is the observed survival time and $\delta_i = I\left(\tilde{y}_i \leq C_i\right)$ with I(.) denoting the indicator function. Note that $\tilde{y}_i = y_i$ if and only if $\delta_i = 1$ (i.e., the subject is uncensored). To deal with the right censoring in the MBPLS algorithm, we use the reweighting method (Datta, 2005) to construct the so-called *adjusted observed survival time* using inverse censoring probability weighting. To briefly summarize the reweighting procedure, let S_C be the survival function of the censoring variable C, i.e., $S_C(t) = P(C > t)$, $t \geq 0$. Then,

$$E\left(\frac{\delta_i y_i}{S_C\left(y_i-\right)}\right) = E\left(E\left(\frac{\delta_i \tilde{y}_i}{S_C\left(\tilde{y}_i-\right)}|\tilde{y}_i\right)\right)$$
$$= E\left(\frac{\tilde{y}_i}{S_C\left(\tilde{y}_i-\right)} E(I(C_i \geq \tilde{y}_i)|\tilde{y}_i)\right) = E(\tilde{y}_i)$$

This identity can be taken as the theoretical basis for estimating the mean observed survival time by the (weighted) sample average $\mu = \left(\sum_{i=1}^{n} \tilde{w}_i y_i\right)/n$, where the weights are $\tilde{w}_i = 0$ for a censored survival time and $\tilde{w}_i = \left(\hat{S}_C\left(y_i-\right)\right)^{-1}$ for an observed survival time. Here, S_C will be calculated using the intercept-only Cox-PH model (Kaplan-Meier) with the status indicator $1 - \delta_i$ (using *survfit* R function)[1].

Partial Least Squares Regression

Partial least squares (PLS) regression, originally introduced by Wold et al. (1983), has been applied in various ill-conditioned linear regression models as both dimension reduction and inference tool. PLS regression works under the assumption of basic latent-decomposition of both X (a $n \times p$ matrix of covariates) and y (a $n \times 1$ matrix of response variable in a univariate case). In contrast to PCA, PLS uses both X and y to construct the latent components ($T = XW$) by maximizing a successive maximization problem. The objective function to find the weight vectors ($W = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$ is as follows:

$$argmax_w \acute{\mathbf{w}} \acute{X} y \acute{y} X \mathbf{w}, \; for\; k = 1, \ldots, K$$
$$s.t. ||\mathbf{w}||_2 = 1,$$

where k is the number of the latent components (tuning parameter or fixed by user), and W is a $p \times k$ matrix of weights (also called direction vector). The latent component T (a $n \times k$ matrix) is then calculated as $T = XW$.

Integrating Cis-Regulatory Quantitative Effects

Genome-wide quantitative trait loci (QTL) mapping enables the determination of genetic loci affecting other Omics, i.e., transcriptome, proteome, and metabolome. Some Omics markers do not directly contribute to the phenotype and affect the disease through other intermediates-Omics. Therefore, considering these QTL associations in each Omics layer can provide more functional information about disease-associated markers. In fact, a QTL-pair is a pair of different Omics (for instance, SNP-CpG in eQTL) that are (highly) associated regarding their effects on the underlying disease (outcome). Hence involving both elements of a QTL-pair (i.e., SNP and gene in an eQTL-pair) in a regression model will not add much more information than involving only one of these elements (at the cost of degrees of freedom and multi-collinearity). One way around this is to consider one element of a QTL-pair in the regression model (such as SNP in an eQTL-pair), then regress out the effect of this element and replace the second element with the residuals. This way, we are considering unwanted/uncorrelated information besides the QTL information in the regression model, avoiding multicollinearity.

As shown in **Figure 3**, for the eQTL-pairs (SNP-gene pairs), we keep the SNPs and replace the genes with residuals by regression out the effect of the SNPs. For the meQTL-pairs (SNP-CpG pairs), we keep the actual DNA methylations (CpGs) and update the genotypes (SNPs) with residuals by regression out the effect of the CpGs. For the eQTM-pairs (gene-CpG pairs) we keep the actual gene expression and update the methylations (CpGs) with residuals by regression out the effect of the genes. The residuals are obtained using the univariate linear regression models. The details of updating each Omics block using QTL-residual are as follows.

Let $X^{(b)} = X^{(b)}_{n \times p_b}$ denotes the observed data in block b ($b = 1, \ldots, B$), a matrix of p_b features/covariates measured for n samples. For ease of presentation, we consider the following $B = 3$ blocks (mRNA expression, genotypes, DNA methylation). However, the proposed modeling framework can accommodate a larger number of blocks and thus other Omics types (such as proteomics and metabolomics) with minor modifications. Thus, let $X^{(1)}_{n \times p_1}, X^{(2)}_{n \times p_2}, X^{(3)}_{n \times p_3}$ denote blocks of mRNA expression of p_1 genes, genotypes of p_2 SNPs, and DNA methylation of p_3 CpGs for n samples, respectively. Next, we split each $X^{(b)}$ based on prior biological knowledge gleaned from eQTL, meQTL, eQTM information. **The first block** (mRNA expression) is split by utilizing the eQTL information as follows:

$$X^{(1)}_{n \times p_1} = \left[X^{(1)}_{n \times q_1} X^{(1)}_{n \times (p_1 - q_1)}\right] = \left[X^{(1)}_{eQTL} X^{(1)}_{Non-eQTL}\right],$$

where q_1 is the number of the eQTL-genes (e.g., with adjusted P-value < 0.05). Whole blood eQTL data are extracted from

[1] https://www.rdocumentation.org/packages/survival/versions/2.11-4/topics/survfit

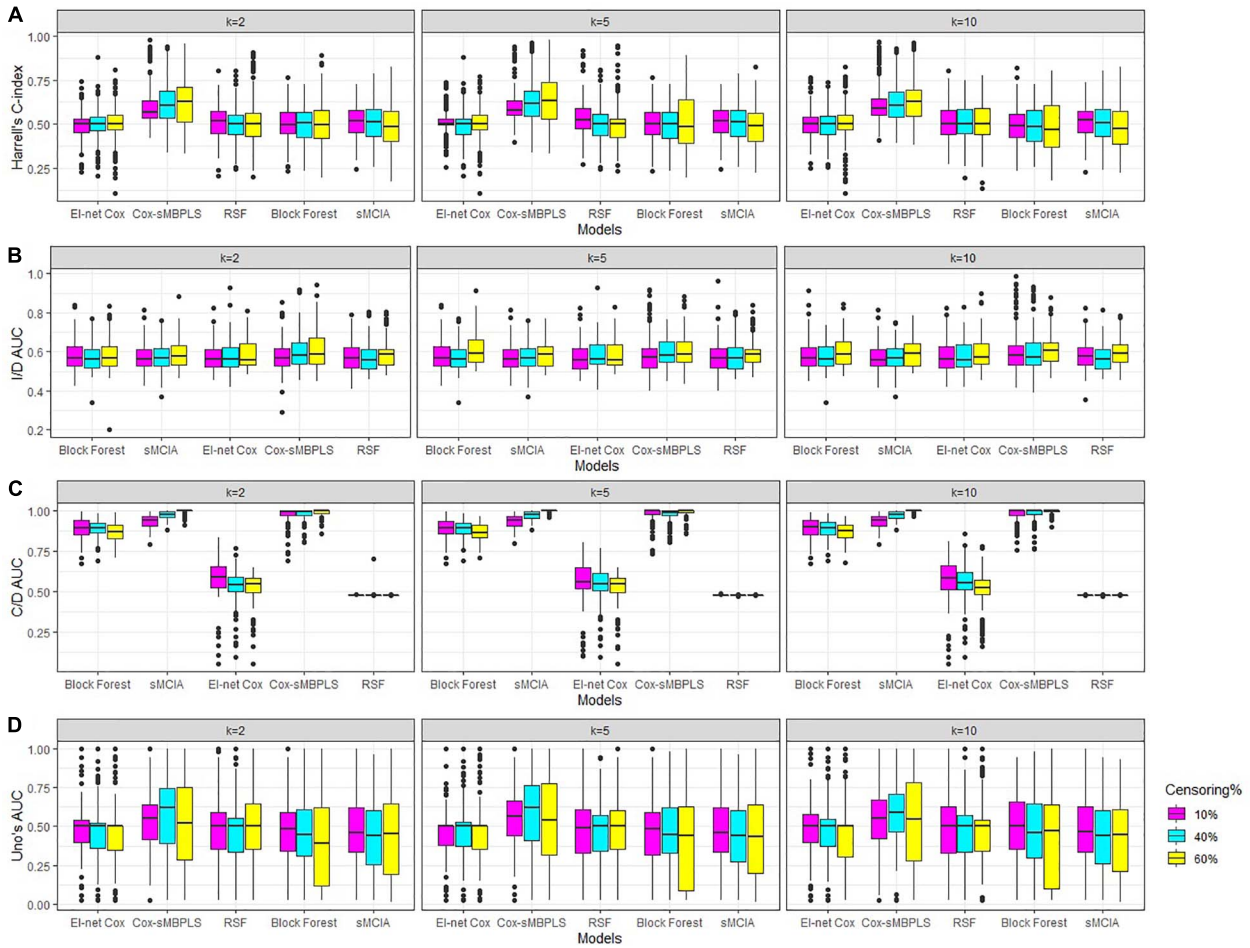


FIGURE 2 | Simulations results for scenarios with a low level of dimensionality. Boxplots for **(A)** Harrell's C-index, **(B)** I/D AUC, **(C)** C/D AUC, and **(D)** Uno's AUC values. Results are shown for different censoring rates ($\delta = 10,\ 40,\ 60\%$) and number of components ($k = 2,\ 5,\ 10$).

the Genotype-Tissue Expression Project (GTEx) (Lonsdale et al., 2013) that are used to extract eQTL-genes and their eQTL-SNP pairs. Specifically, suppose that SNP_j^{eQTL} is a vector of SNPs which are in eQTL with (single) $gene_j$ $(j = 1, \ldots, q_1)$ and $SNP_j^{eQTL} \in X^{(2)}$. By regressing out the effect of these SNPs, the so-called eQTL residuals (u_j^{eQTL}) are defined as the residuals of the regression of the j^{th} eQTL-gene ($\in X_{eQTL}^{(1)}$) on SNP_j^{eQTL}, as follows:

$$gene_j \propto \sum_{j=1}^{q_1} \alpha SNP_j^{eQTL} + u_j^{eQTL},\ \ where\ gene_j \in\ X_{eQTL}^{(1)}$$

We then update $X_{n\times p_1}^{(1)}$ by replacing $X_{eQTL}^{(1)}$ with u_j^{eQTL}. Therefore, the updated $X_{n\times p_1}^{(1)}$ becomes:

$$X_{n\times p_1}^{(1)*} = \left[u_j^{eQTL}\ X_{Non-eQTL}^{(1)}\right].$$

The second block (genotypes) is split by utilizing the meQTL information as follows:

$$X_{n\times p_2}^{(2)} = \left[X_{n\times q_2}^{(2)}\ X_{n\times(p_2-q_2)}^{(2)}\right] = \left[X_{meQTL}^{(2)}\ X_{Non-meQTL}^{(2)}\right],$$

where q_2 is the number of the meQTL-SNPs (with adjusted P-value < 0.05). Whole blood *cis*-meQTL data were extracted from BIOS QTL (Zhernakova et al., 2017) that are used to extract the meQTL-SNPs and their meQTL-CpG pairs. Specifically, suppose that $CpG_{\acute{j}}^{meQTL}$ is a vector of CpGs which are in meQTL with (single) $SNP_{\acute{j}}$ $(\acute{j} = 1, \ldots, q_2)$ and $CpG_{\acute{j}}^{meQTL} \in X^{(3)}$. By regressing out the effect of these CpGs, the so-called "meQTL residuals" ($u_{\acute{j}}^{meQTL}$) are defined as the residuals of the regression of $\acute{j}^{th}$ meQTL-SNP ($\in X_{meQTL}^{(2)}$) on $CpG_{\acute{j}}^{meQTL}$, as follows:

$$snp_{\acute{j}} \propto \sum_{\acute{j}=1}^{q_2} \acute{\alpha} cpg_{\acute{j}}^{meQTL} + u_{\acute{j}}^{meQTL},\ \ where\ snp_{\acute{j}} \in X_{meQTL}^{(2)}.$$

We then update $X^{(2)}_{n\times p_2}$ by replacing $X^{(2)}_{meQTL}$with u^{meQTL}_{j}. Therefore, updated $X^{(2)}_{n\times p_2}$becomes:

$$X^{(2)*}_{n\times p_2} = \left[u^{meQTL}_{j}\ X^{(2)}_{Non-meQTL}\right].$$

The third block (DNA methylations) is split by utilizing the eQTM information as follows:

$$X^{(3)}_{n\times p_3} = \left[X^{(3)}_{n\times q_3}\ X^{(3)}_{n\times(p_3-q_3)}\right] = \left[X^{(3)}_{eQTM}\ X^{(3)}_{Non-eQTM}\right],$$

where q_3 is the number of the eQTM-CpGs (with adjusted P-value < 0.05). Whole blood *cis*-eQTM data were extracted from BIOS QTL (Zhernakova et al., 2017) that are used to extract the eQTM-CpGs and their eQTm-gene pairs. Specifically, suppose that $gene^{eQTM}_{j"}$ is a vector of genes which are in eQTM with (single) CpG $j"$ ($j" = 1, \ldots, q_3$). By regressing out the effect of these genes, the so-called "eQTM residuals" ($u^{eQTM}_{j"}$) are defined as the residuals of the regression of the $j"^{th}$ eQTM-CpG ($\in X^{(3)}_{eQTM}$) on $gene^{eQTM}_{j"}$, as follows:

$$cpg_{j"} \propto \sum_{j"=1}^{q_3} \alpha" gene^{eQTM}_{j"} u^{eQTM}_{j"},\ where\ cpg_{j"} \in X^{(3)}_{eQTM}.$$

We then update $X^{(3)}_{n\times p_3}$ by replacing $X^{(3)}_{eQTM}$ with $u^{eQTM}_{j"}$. Therefore, updated $X^{(3)}_{n\times p_3}$becomes:

$$X^{(3)*}_{n\times p_3} = \left[u^{eQTM}_{j"} X^{(3)}_{Non-eQTM}\right].$$

The illustration of the multi-Omics data structure and Omics-block updating procedure is presented in **Figure 1**.

Supervised Cox-sMBPLS Algorithm

Let $X^{(b)}$, $b = 1, 2, 3$ ($X^{(b)} = X^{(b)*}$ as explained in section "Integrating Cis-regulatory Quantitative Effects") and $\mathbf{y}$ ($\mathbf{y} = \mathbf{y}^*$ as explained in section "Reweighted Survival Time") be the covariate matrices (blocks) and response vector on the same n samples, respectively. In each block, the dimensionality of the block can be reduced by taking a linear combination of the covariates $\tau^{(b)} = X^{(b)}\mathbf{w}^{(b)}$, where $\mathbf{w}^{(b)}$ is direction vector (also called weight vector) that express the importance of each covariate on the latent component $\tau^{(b)}$ (a $n \times 1$ matrix). Li et al. (2012), suggested using a weighted sum of the latent components over the blocks as the combined latent component. Therefore, we define $\tau = \sum_{b=1}^{3} \tau^{(b)}\omega^{(b)}$ as the combined latent component, where $\omega^{(b)}$ ($\omega^{(b)} > 0$) is the weight for block b, which indicates the contribution of this block to the covariance structure of the input and response ($\mathbf{y}$) data. We can then posit the following optimization problem for calculating the latent components across all blocks:

$$\max_{\mathbf{w}^{(b)}} cov(\tau, \mathbf{y}),\ with\ \tau = \sum_{b=1}^{3} \tau^{(b)}\omega^{(b)},\ and \tau^{(b)} = X^{(b)}\mathbf{w}^{(b)},$$
$$\omega^{(b)} \in \mathbb{R}^{+}, X^{(b)} \in \mathbb{R}^{n\times p_b}, \mathbf{w}^{(b)} \in \mathbb{R}^{p_b}, \mathbf{y} \in \mathbb{R}^{n},$$
$$subject\ to\ ||\mathbf{w}^{(b)}||_2 = ||\omega||_2 = 1. \quad (1)$$

A straightforward extension of problem (1) to the sparse version could be obtained by adding an L_1 penalty on direction vector $\mathbf{w}^{(b)}$; i.e., $||\mathbf{w}^{(b)}||_1 \geq \lambda$, for some positive tuning parameter λ. However, Jolliffe et al. (2003), showed, via an example, that this formulation may not lead to a sufficiently sparse solution. The sparsity issue for PCA was first considered by Zou et al. (2006) by imposing both L_1&L_2 constraints on the weight coefficients.

In the case of PLS, Chun and Keleş (2010) used the same approach by imposing an L_1 constraint onto a surrogate of the direction vector. Therefore, a generalized version of the optimization problem using a combined L_1&L_2 regularization component becomes:

$$\min_{\mathbf{w}^{(b)}, c^{(b)}} \left\{-\kappa\omega^{(b)^2}\acute{\mathbf{w}}^{(b)}Z^{(b)}\mathbf{w}^{(b)} + \omega^{(b)^2}(1-\kappa)\left(c^{(b)} - \mathbf{w}^{(b)}\right)' Z^{(b)}\right.$$
$$\left.\left(c^{(b)} - \mathbf{w}^{(b)}\right) + \lambda_1||c^{(b)}||_1 + \lambda_2||c^{(b)}||_2\right\},$$
$$subject\ to\ ||\mathbf{w}^{(b)}||_2 = ||\omega||_2 = 1. \quad (2)$$

For any non-negative $\lambda_1$and λ_2. $Z^{(b)} = \acute{X}^{(b)}\mathbf{y}\acute{\mathbf{y}}X^{(b)}$, $c^{(b)}$ is the surrogate of the direction vector in block b, which is kept close to the original direction vector $\mathbf{w}$; and κ is a tuning parameter. κ is the concave-penalty parameter to control the amount of the weight is given to the concave part of the objective function [$\acute{\mathbf{w}}^{(b)}Z^{(b)}\mathbf{w}^{(b)}$], and therefore, to control the local-solution issue. For more details and history of recasting from equation (1) to equation (2), see **Supplementary Section 1.1**.

Bastien et al. (2015) also employed this method to propose a sparse PLS for censored data. Chun and Keleş (2010) solved problem (2) by alternatively iterating between solving for $\mathbf{w}$ after fixing c and solving for c for fixed $\mathbf{w}$. However, they showed that in the case of a univariate response, problem (2) does not depend on and often needs a large λ_2 value to be solved. Therefore, we use $\lambda_2 \to \infty$ which yields the special case of the elastic-net regularization, called univariate soft-thresholding (Zou and Hastie, 2005). Hence, the solution to problem (2) has the following closed form:

$$\hat{c}^{(b)} = \left(\frac{\acute{X^{(b)}}\mathbf{y}}{||\acute{X^{(b)}}\mathbf{y}||} - \frac{\lambda}{2}\right)_{+} sign\left(\frac{\acute{X^{(b)}}\mathbf{y}}{||\acute{X^{(b)}}\mathbf{y}||}\right), \lambda \geq 0 \quad (3)$$

where $\frac{\acute{X^{(b)}}\mathbf{y}}{||\acute{X^{(b)}}\mathbf{y}||}$ is the first direction vector, and $soft-threshold(\Phi, \lambda) = \left(\Phi - \frac{\lambda}{2}\right) sign(\Phi)$ is the soft-thresholding operator with a fixed non-negative parameter λ ($\lambda = \lambda_1$). The algorithm is then followed by a PLS regression on the selected variables, and iterated with updating the response variable, $\mathbf{y}$. The proof for the solution in equation (3) is

Algorithm 1 | Supervised Cox-sMBPLS algorithm.

1. Calculating reweighted survival time (y*): Calculate the reweighted survival time as $y_i^* = \frac{y_i.\delta_i}{\hat{S}_C}$, where $\hat{S}_C$ is the estimated survival function for censoring variable (C) and y_i is time to a specific event for the i^{th} sample, as explained in section "Reweighted Survival Time".

2. Incorporating the *cis*-regulatory quantitative effects (eQTL, eQTM, meQTL) information: Split and update each block employing the corresponding *cis*-regulatory quantitative effects. This process is fully explained in section "Integrating Cis-regulatory Quantitative Effects."

3. Computing the latent components and block-weights: by applying the sMBPLS algorithm on $X^{(b)}$ (as independent variables) and y^* (as dependent variable) as follows:

do for $k = 1, \ldots, K$ where k (a tuning parameter) is the number of the latent components,

3.1 Set $\widehat{\beta}_{PLS}^{(b)} = 0$, $\Omega^{(b)} = \{\ \}$, $\mathbf{y} = \mathbf{y}^*$ and $k = 1$, where $b = 1, 2, 3$.

3.2 $\mathbf{w}^{(b)} = soft - threshold\left(\frac{\acute{X}^{(b)*}\mathbf{y}^*}{||\acute{X}^{(b)*}\mathbf{y}^*||_2}, \lambda\right)$, where λ is a non-negative sparsity parameter.

3.3 $\tau^{(b)} = X^{(b)}\mathbf{w}^{(b)}$, $b = 1, 2, 3$

3.4 $\tau = [\tau^{(1)}, \tau^{(2)}, \tau^{(3)}]$

3.5 $\omega = \acute{\mathbf{y}}^*\tau$

3.6 $\mathbf{T} = \tau\acute{\omega}$

3.7 Update $\mathbf{\Omega}^{(b)}$ as $\left\{j^{(b)} : \hat{w}_{j^{(b)}}^{(b)} \neq 0\right\} \cup \left\{j^{(b)} : \widehat{\beta}_{PLS_{j^{(b)}}}^{(b)} \neq 0\right\}$, where $j^{(b)} \in \{1, \ldots, p_b\}$

3.8 Fit PLS regression to $X^{(b)*}$ (updated based on $\mathbf{\Omega}^{(b)}$) and $\mathbf{y}^*$ in each block ($b = 1, 2, 3$) using k number of latent components.

3.9 Update $\widehat{\beta}_{PLS}^{(b)}$ with new estimated coefficients resulted from the PLS regression in step 3.8. Update $\mathbf{y}^*$ as $\mathbf{y}^* \rightarrow \mathbf{y}^* - X^{(b)*}\widehat{\beta}_{PLS}^{(b)}$

4. Final Cox-sMBPLS model: Fit a Cox-PH model with (y_i, δ_i) and remaining latent components from the sMBPLS algorithm.

provided in **Supplementary Section 1.1**. The full algorithm is described below.

The conjugacy of direction vectors (similar to orthogonality issue in PCA-kind problems) is addressed by keeping the Krylov subsequence structure of the direction vectors in a restricted X-space of selected variables ($X^{(b)} \in \mathbf{\Omega}^{(b)}$) (Chun and Keleş, 2010). Specifically, at each step of the **Algorithm 1**, it searches for relevant variables, the so-called active variables (updated in step 3.7), by optimizing equation (2) and updates all direction vectors to form a Krylov subsequence on the subspace of the active variables. This is simply achieved by conducting PLS regression by using the selected variables (see step 3.8, **Algorithm 1**).

Initial values are set in step 3.1 and sparse weight (direction) vectors are calculated in step 3.2 where λ is a non-negative (sparsity) tuning parameter which is tuned using a k-fold cross-validation (see section "Parameter Tuning and Model Performance Evaluation" for details). In step 3.3 the latent components for each block are calculated ($\tau_{n\times1}^{(1)}, \tau_{n\times1}^{(2)}, \tau_{n\times1}^{(3)}$) which are then combined using a weighted sum over the blocks to calculate the combined latent component ($\mathbf{T}_{n\times1}$) in step 3.6. Blocks' weight (ω) are calculated in Step 3.5. The so-called active variables set ($\mathbf{\Omega}^{(b)}$) is then updated in step 3.7 followed by a PLS regression using active variables $X^{*(b)} \in \mathbf{\Omega}^{(b)}$ as covariates and $\mathbf{y}^*$ as response variable. The PLS regression is fit using *wpls* R function, adapted from *spls* R package[2]. Response variable ($\mathbf{y}^*$) is then updated in step 3.9. Note that $\acute{X^{(b)}}*$ is scaled, including genotype data (categorical variable), as suggested by Tibshirani (1997). To scale the categorical genotypes $[\acute{X^{(2)}}*]$, we considered the fact that encoded genotypes (0, 1, 2) are quantitative measures correspond to the number of minor alleles in the genotype. Latent components are computed using the updated data, $X^{(b)*}$ and $\mathbf{y}^*$ (step 3.9). The solution to the optimization function (2) also enables us to identify multi-Omics modules. These modules are linear combinations of multiple Omics profiles with large absolute values of $\mathbf{w}^{(b)}$ if happen together. It is possible to apply different sparsity parameters for each block and or direction vector, which is avoided here due to the computational burden of tuning multiple parameters. The illustration of the data structure and the supervised Cox-sMBPLS algorithm is provided in **Figure 1**.

Parameter Tuning and Model Performance Evaluation

Cross-validation (CV) is used to tune the number of components (k) and the sparsity (λ) hyper-parameters that lead to the best prediction performance. In principle, we can try different combinations of k (number of the latent components) and λ (sparsity). The chosen k and λ are the ones giving the highest model performance measures. For the real data analysis, we considered $1 = k = min\left\{p, \frac{\nu-1}{\nu}n\right\}$ where p is the total number of the covariates, ν is the fold number in the (k-fold) CV, and n is the sample size. This upper bound for the number of the latent components is suggested by Chung et al. (2012). In each iteration, the supervised Cox-sMBPLS model is trained using a training-set (in the numerical work, we set it to 80% of data). The test data-set (remaining 20% of data) is then used to evaluate the predictive performance using Harrell's C-index (Harrell et al., 1982) ($\frac{\sum_{ij\ i}\ I(\hat{y}_i=\hat{y}_j)I(y_i=y_j)}{\sum_{ij\ i}I(y_i=y_j)}$) and time-specific area under the ROC-curve, AUC. We used the incident/dynamic (I/D) ROC-curves (Heagerty and Zheng, 2005) and Uno (Uno et al., 2007)

[2]https://cran.r-project.org/web/packages/spls/index.html

TABLE 1 | Simulation settings based on the different number of latent components (k), censoring rate (δ) and p_b number of features in block b.

Scenario #	Censoring	Dimensionality	Number of Components
1**	$\delta = 10\%$ (Low)	$p_b = 100 = \upsilon^{(b)*}$ (Low)	$k = 2,\ 5, 10$
2		$p_b = 1000 = \upsilon^{(b)}$ (Moderate)	$k = 2,\ 5, 10$
3		$p_b = 10\,000 = \upsilon^{(b)}$ (High)	$k = 2,\ 5, 10$
4	$\delta = 40\%$ (Moderate)	$p_b = 100 = \upsilon^{(b)*}$ (Low)	$k = 2,\ 5, 10$
5		$p_b = 1000 = \upsilon^{(b)}$ (Moderate)	$k = 2,\ 5, 10$
6		$p_b = 10\,000 = \upsilon^{(b)}$ (High)	$k = 2,\ 5, 10$
7	$\delta = 60\%$ (High)	$p_b = 100 = \upsilon^{(b)*}$ (Low)	$k = 2,\ 5, 10$
8		$p_b = 1000 = \upsilon^{(b)}$ (Moderate)	$k = 2,\ 5, 10$
9		$p_b = 10\,000 = \upsilon^{(b)}$ (High)	$k = 2,\ 5, 10$

Three -Omics blocks, b = 3, (mRNA expression, genotypes and DNA methylation) sampled from the same n = 91 samples are considered.
** We define $\upsilon^{(b)}$, as the weight of each block, relative to the total number of genes: $\upsilon^{(1)} = \frac{27\,645}{27\,645} = 1$, $\upsilon^{(2)} = \frac{578\,846}{27\,645} = 20.9$, $\upsilon^{(3)} = \frac{12\,283}{27\,645} = 0.4$.*
*** 1350 independent replicates for each scenario, with 450 replicates for each k (k = 2,5,10) in scenarios with low and moderate levels of dimensionality, and 300 replicates for each k in the scenario with a high level of dimensionality.*

and Chambless (Chambless and Diao, 2006) estimators of cumulative/dynamic (C/D) AUC (more information is provided in **Supplementary Section 1.2**).

Data Source

The analyzed data set contains information on 91 subjects with heart failure (HF); namely, with preserved ejection fraction (HFpEF) and reduced ejection fraction (HFrEF). Further, 47% of them experienced death or a hospitalization event. The original discovery cohort included 103 HF (HFpEF and HFrEF) patients with complete data of all Omics types (mRNA expression, genotypes, and DNA methylation), 12 of which were removed due to sex mismatch and $n = 91$ patients remained in the analysis. The subjects were recruited from cardiology clinics during a four-year period (2011–2015) at the University of Illinois at Chicago (UIC). All patients provided written, informed consent (Mansour et al., 2016; Duarte et al., 2018).

RNA profiles were obtained by using the Affymetrix Human Gene 2.0 ST array. After quality control procedures, 27 645 genes were kept for subsequent analysis. Genotypes were measured by high-density genome-wide bead array genotyping (Affymetrix Axiom PanAfrican Array). We excluded SNPs with a missing rate $\geq$10%, monomorphic SNPs with MAF $< 0.01\%$ and SNPs on the negative strands. Additional genotypes were imputed based on a two-step approach. First, the samples were phased into a series of estimated haplotypes, and then, imputation was performed on them. After imputation, genotypes with $R^2 =$ 80% were excluded to keep only high-quality imputed profiles. We then performed linkage disequilibrium pruning (LDP). Thereafter, whole blood *cis*-eQTL data from the Genotype-Tissue Expression Project (GTEx) (Lonsdale et al., 2013) were used to remove non-eQTL-SNPs with adjusted eQTL-pvalue > 0.1. We applied this filter as part of the pre-analysis feature selection procedure since it is shown that GWAS eQTL-SNPs tended to be more significant compared to non-eQTL-SNPs (Gorlov et al., 2019). In total, 578 846 SNPs remained in the study. DNA methylation profiles were measured using the Illumina Infinium Human Methylation 450 (450K) BeadChip array. Whole blood *cis*-eQTM data from BIOS QTL (Zhernakova et al., 2017) were then utilized to remove the non-eQTM-CpGs with eQTM-pvalue > 0.1. In total, 12 283 CpGs remained in the study.

RESULTS

Results on Simulated Data

We performed a set of simulation studies in order to evaluate the prediction accuracy of the proposed supervised Cox-sMBPLS model. The settings under consideration aim to control the redundancy within the Omics profiles (via a soft-threshold), the association between the Omics profiles (via *cis*-regulatory quantitative effects), and the relevance of each Omics profile or a combination of them to explain the survival probabilities. We compared the proposed model to elastic-net Cox-PH (El-net Cox) (Simon et al., 2011), random survival forest (RSF) (Ishwaran et al., 2008), Block Forest (Hornung and Wright, 2019), and multiple co-inertia analysis (MCIA) (Min and Long, 2020). All models are trained on 80% of the samples and tested on the left-out 20% portion. Further, cross-validation (CV) is used to tune the hyperparameters for all methods considered. To generate more realistic samples of the Omics profiles, we randomly sampled from a real-world multi-Omics data set, which is described in the next section. We simulated 10 800 replicates based on combinations of the following factors for a total of 9 scenarios: the number of the latent components k, the censoring rate δ the number of blocks b, and the number of features p_b in block b. Details on the simulation settings are provided in **Table 1**.

We sample true predictor matrices X_T^b $(b = 1, 2, 3)$ of dimension $n = p_b$, with fixed sample size $n = 91$. Matrices X_T^b $(b = 1, 2, 3)$ are random samples (without replacement) from the real Omics data presented in the next section. For the construction of true latent components (τ_T), we assume that some of the features in each block have small or no effect on the response variable by specifying sparse (true) direction vector ($\mathbf{w}_T$). These weights are sampled from the distribution of the weights collected from a standard sparse PLS on a random sample of the Omics data. Therefore, true latent components are sparse across all simulations. The response variable (y_i, δ_i) for sample i $(i = 1, , 91)$ is simulated using a flexible-hazard

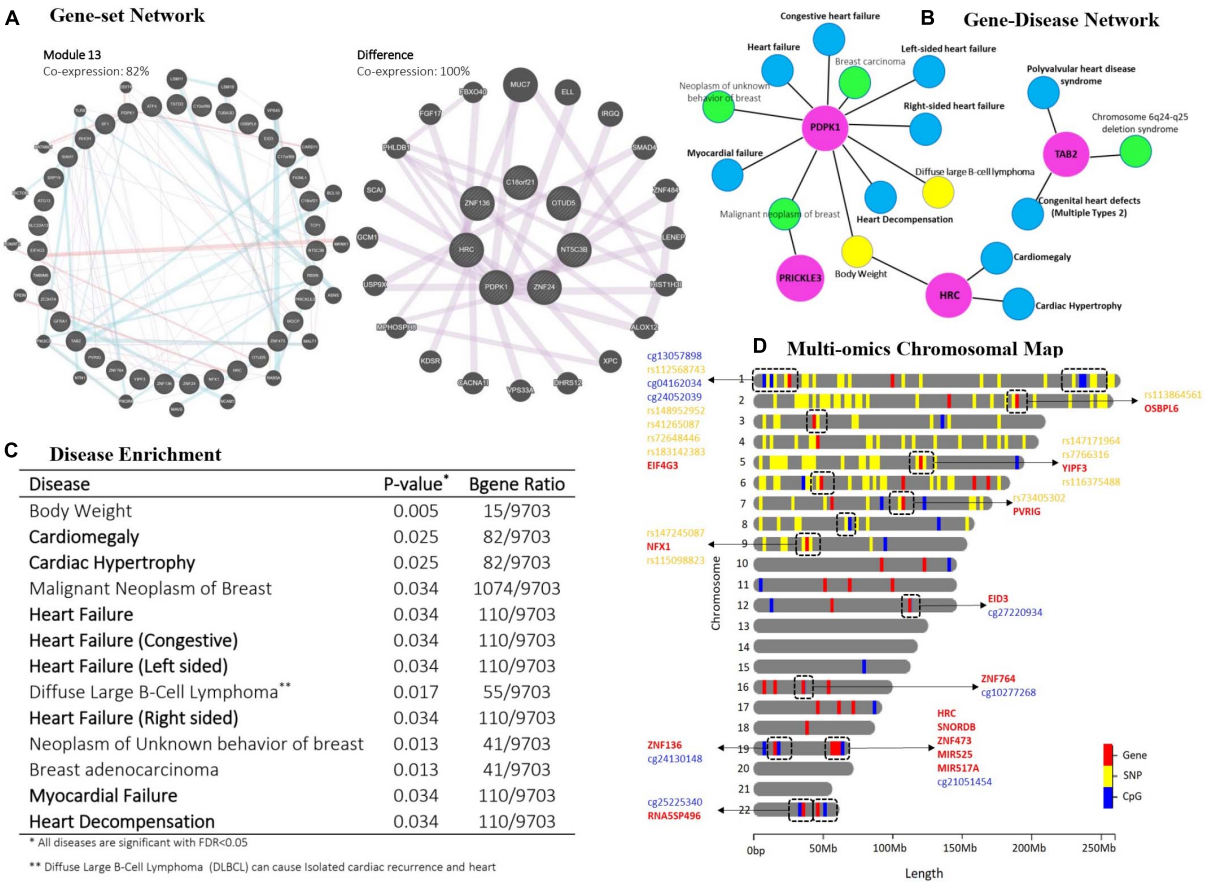
model (Harden and Kropko, 2019). We simulate the replicates using twenty different seed numbers and the results assess the stability of the models against the seed numbers. More details, including the simulation algorithm are provided in **Supplementary Section 2**. The results of the simulation studies for the low level of dimensionality are shown in **Table 2** and **Figure 2** (for $k = 2, 5$). See **Supplementary Table 1** for the full results of the low level of dimensionality (including $k = 10$). The results of the moderate and high levels of dimensionality are presented in **Supplementary Tables 2, 3**) and are very similar to the results of the low level of dimensionality.

The proposed supervised Cox-sMBPLS model exhibited better performance than El-net Cox, RSF, Block forest and MCIA models in both scenarios with low and moderate level of dimensionality. The prediction performance (C-index) and feature-selection performance (AUCs) of the proposed Cox-sMBPLS model remained higher than other models regardless of the different changing parameters. When increasing the censoring rate from 10 to 60%, the feature-selection performance (C/D AUC) of all models decreased (except for MCIA) (**Table 2**): in Cox-sMBPLS decreased by 2% for $k = 2$, 4% for $k = 5$ and $k = 10$; in El-net Cox decreased by 8%, 7%, and 11% for $k = 2, 5, 10$, respectively; in RSF decreased by 11%, 3%, and 10% for $k = 2, 5, 10$, respectively; in Block Forest decreased by 6%, 5%, and 5% for $k = 2, 5, 10$, respectively; in MCIA decreased by 3% and 0% for $k = 2, 5$, respectively, and increased by 2% for $k = 10$. Amongst these models, the proposed Cox-sMBPLS (2–4% decrease) and Block Forest (5–6% decrease) are more stable than El-net Cox (7-11% decrease), RSF (3–11% decrease), and MCIA (0–3% decrease and/or 2% increase) against the censoring rate. The feature selection results showed less stability against censoring rate using other performance measures (I/D AUC, and Uno's AUC). I/D AUC, and Uno's AUC tended to increase while increasing the censoring rate. This may be due to the relatively low number of events and the following low number of patient-pairs used to estimate the measures. Rahman et al. (2017) experienced the same results in an extensive simulation study comparing the different performance measures for survival models. In general, the results were more stable against censoring rate by increasing the number of the latent components, k (see **Supplementary Tables 1–3**). In most of the settings, the variability (SDs) of all measures increased by increasing the censoring rate, as expected. Similar results for other settings are presented in **Supplementary Tables 2, 3** and **Supplementary Figures 1, 2**. When increasing the dimensionality of the predictors, the performance of all models decreased, even though the Cox-sMBPLS model continued to outperform the other ones. Based on the AUC values, Cox-sMBPLS exhibited a superior performance to El-net Cox, RSF, Block Forest, and MCIA, having higher probability of selecting the correct (important) features (**Figure 2**). In **Table 2**, the range of the prediction performance (C-index) of the proposed Cox-sMBPLS model was 0.60–0.64, for El-net Cox was 0.49–0.51, for RSF was 0.50–0.53, for Block Forest was 0.49–0.51, and for MCIA was 0.48-0.51 for different changing parameters. There is a recent benchmark analysis (Jardillier et al., 2020) of lasso-like penalties (including ridge, lasso, adaptive lasso, and elastic-net) of the Cox model where the authors showed

TABLE 2 | Simulation results for the scenarios with a low level of dimensionality (total of 2230 features, and $n = 91$).

Censoring %	Measure	Number of components									
		k = 2					k = 5				
		Cox-sMBPLS	El-net Cox	RSF	Block forest	MCIA	Cox-sMBPLS	El-net Cox	RSF	Block forest	MCIA
10% (Low)	C-index	**0.60** (0.10)	0.49 (0.08)	0.50 (0.10)	0.49 (0.09)	0.51 (0.09)	**0.60** (0.09)	0.51 (0.07)	0.53 (0.10)	0.51 (0.10)	0.51 (0.09)
	C/D AUC*	**0.95** (0.13)	0.38 (0.30)	0.46 (0.06)	0.87 (0.13)	0.91 (0.12)	**0.97** (0.09)	0.36 (0.29)	0.46 (0.09)	0.87 (0.12)	0.91 (0.12)
	I/D AUC**	**0.58** (0.07)	0.57 (0.06)	**0.58** (0.07)	0.57 (0.07)	0.57 (0.07)	**0.58** (0.08)	0.57 (0.07)	0.57 (0.08)	0.57 (0.07)	0.57 (0.07)
	Uno's AUC***	**0.52** (0.22)	0.46 (0.18)	0.48 (0.22)	0.46 (0.23)	0.49 (0.23)	**0.53** (0.23)	0.44 (0.18)	0.46 (0.23)	0.46 (0.23)	0.45 (0.23)
40% (Moderate)	C-index	**0.62** (0.11)	0.49 (0.09)	0.51 (0.10)	0.49 (0.10)	0.51 (0.11)	**0.63** (0.11)	0.49 (0.09)	0.51 (0.10)	0.49 (0.11)	0.51 (0.11)
	C/D AUC	**0.94** (0.18)	0.35 (0.27)	0.39 (0.19)	0.84 (0.19)	0.90 (0.20)	**0.94** (0.17)	0.36 (0.27)	0.37 (0.19)	0.84 (0.19)	0.90 (0.20)
	I/D AUC	**0.60** (0.08)	0.57 (0.07)	0.57 (0.08)	0.56 (0.07)	0.57 (0.07)	**0.60** (0.07)	0.58 (0.07)	0.58 (0.08)	0.57 (0.07)	0.57 (0.07)
	Uno's AUC	**0.54** (0.29)	0.44 (0.22)	0.45 (0.23)	0.44 (0.25)	0.42 (0.24)	**0.56** (0.28)	0.45 (0.22)	046 (0.23)	0.45 (0.25)	0.42 (0.24)
60% (High)	C-index	**0.63** (0.13)	0.50 (0.10)	0.51 (0.13)	0.49 (0.14)	0.48 (0.13)	**0.64** (0.12)	0.50 (0.10)	0.50 (0.12)	0.50 (0.16)	0.48 (0.12)
	C/D AUC	**0.93** (0.23)	0.30 (0.27)	0.35 (0.21)	0.81 (0.20)	**0.93** (0.24)	**0.93** (0.23)	0.29 (0.26)	0.34 (0.21)	0.82 (0.20)	**0.93** (0.24)
	I/D AUC	**0.61** (0.09)	0.58 (0.07)	0.59 (0.06)	0.57 (0.09)	0.58 (0.07)	0.60 (0.08)	0.59 (0.07)	0.59 (0.06)	**0.61** (0.09)	0.58 (0.07)
	Uno's AUC	**0.50** (0.31)	0.45 (0.24)	0.47 (0.27)	0.40 (0.28)	0.41 (0.29)	**0.51** (0.32)	0.45 (0.24)	0.46 (0.27)	0.40 (0.30)	0.41 (0.29)

Performance measures are averaged over S = 450 simulations. SDs are shown in the parentheses.
** Chambless estimator of cumulative/dynamic (C/D) AUC.*
*** Incident/dynamic (I/D) AUC.*
**** Uno estimator of cumulative/dynamic (C/D) AUC.*
Bold values show the highest performance for each measurement and k.



Disease	P-value*	Bgene Ratio
Body Weight	0.005	15/9703
Cardiomegaly	0.025	82/9703
Cardiac Hypertrophy	0.025	82/9703
Malignant Neoplasm of Breast	0.034	1074/9703
Heart Failure	0.034	110/9703
Heart Failure (Congestive)	0.034	110/9703
Heart Failure (Left sided)	0.034	110/9703
Diffuse Large B-Cell Lymphoma**	0.017	55/9703
Heart Failure (Right sided)	0.034	110/9703
Neoplasm of Unknown behavior of breast	0.013	41/9703
Breast adenocarcinoma	0.013	41/9703
Myocardial Failure	0.034	110/9703
Heart Decompensation	0.034	110/9703

* All diseases are significant with FDR<0.05

** Diffuse Large B-Cell Lymphoma (DLBCL) can cause Isolated cardiac recurrence and heart

FIGURE 3 | Post-hoc analyzes for multi-mics modules 10 and 13 resulted from the Cox-sMBPLS model. **(A)** gene-set network analysis. **(B)** disease ontology for *PDPK1*, *TAB2*, *PRICKLE3*, and *HRC* genes which are presented in both modules. **(C)** disease enrichment analysis. **(D)** chromosomal map. Dashed boxes show the Omics profiles, which are located in a small window on the same chromosome.

that the lasso-like penalized Cox model (including El-net Cox) have the potential of having ≥50% of false discovery proportion. This is consistent with our simulation results where El-net Cox performs poorly in most of the scenarios.

Overall, the proposed supervised Cox-sMBPLS method outperformed all competing methods (El-net Cox, RSF, Block Forest, and MCIA) regarding the exact survival prediction and feature-selection power. Moreover, this method showed less sensitivity to the selection of the tuning parameters and censoring rate compared to competing methods.

Results on a Heart-Failure Discovery Cohort

The supervised Cox-sMBPLS model (on HF cohort) retained $k = 15$ multi-Omics modules (i.e., a combination of genes, SNPs, and CpGs affecting the survival probability when occurring together). k is tuned using a 5-fold CV within the range of $1 = k = 73$. The upper boundary ($k = 73$) is calculated based on Chung et al. (2012) suggestion as $min\left\{p, \frac{\nu-1}{\nu}n\right\}$, where p is the total number of the covariates, ν is the fold number in the (k-fold) CV, and n is the sample size. Results are provided in **Supplementary Section 2.3**. **Figure 3** presents two significant multi-Omics modules (10, 13) in the final Cox model (p-value < 0.1) in detail. Module 10 contains 375 features (33 genes, 308 SNPs, 34 CpGs) and module 13 contains 497 features (49 genes, 399 SNPs, 49 CpGs). There are 122 Omics profiles (16 genes, 91 SNPs, and 15 CpGs), included in module 13 but not module 10. Details for these 122 features are also included in **Figure 3** (see also **Supplementary Tables 4–6**).

We also compared the prediction performance of the Cox-sMBPLS model to El-net Cox (Simon et al., 2011), RSF (Ishwaran et al., 2008), Block Forest (Hornung and Wright, 2019) and MCIA (Min and Long, 2020) (see **Supplementary Table 7** for prediction performance measures). All models are trained on 80% of the samples and tested on the left-out 20% portion. The proposed supervised Cox-sMBPLS showed better performance regarding both C-index and AUC measures.

To interpret the biological relevance of the significant multi-Omics modules enrichment analysis of their Omics profiles using network-based resources and disease ontology is undertaken. Specifically, we performed a gene-set network analysis (**Figure 3A**) using GeneMANIA (Warde-Farley et al., 2010), gene-disease network (**Figure 3B**), and disease enrichment

analysis (**Figures 2D**, **3C**), both using DisGeNET knowledge platform (Piñero et al., 2020). Gene-set network analysis shows 82% and 100% co-expression between the genes in multi-Omics module 13 and the difference between modules 10 and 13, respectively. Modules 10 (p = 0.097) and 13 (p = 0.059) are the two significant modules (see **Supplementary Table 4**). There is a 61% decrease in P-value from module 10 to module 13 (from 0.097 to 0.059). To figure out the legitimacy of this strengthening in the resulted association (from module 10 to module 13), we removed the overlaps between these two modules and ran a gene-set network analysis for the remaining genes (which are causing this 61% decrease in P-value). The result showed 100% co-expression between them, which proves that this boost from module 10 to module 13 is biologically genuine and worthy to run further functional validations to study them for finding novel biomarkers. Moreover, gene-disease network analysis of the selected genes common in both modules (*PDPK1*, *TAB2*, *PRICKLE3*, and *HRC*) also confirms the role of these genes in heart complications. Disease enrichment results similarly show that the genes in the multi-Omics modules are mainly enriched for heart disease, such as heart failure, cardiac hypertrophy, and myocardial failure. A brief discussion of the common genes follows.

PDPK1 (Phosphoinositide-dependent protein kinase-1) is the *PDK1* protein coding gene and also a part of the *AGC* super family of protein kinases which have been well documented for playing a crucial role in heart complications (Marrocco et al., 2019). It has also been reported as a component of the TGF-β/smad signaling pathway which leads to decompensation and heart failure (Kuzmanov et al., 2016). Histidine-rich calcium binding protein (*HRC*) can affect Ca2+ cycling in the sarcoplasmic reticulum (SR) that could cause the mitochondrial death pathway and enhance cardiac function in failure heart (Park et al., 2012). *TAB2* (*TAK1* binding protein-2) is known to play an important role in cardiac development and has recently received more attention in heart diseases. There has been recent research suggesting *TAB2* and its signaling network (*TAB2-TAK1*) as novel therapeutic targets in heart complications (Yin et al., 2017; Cheng et al., 2020). Moreover, a first report of a Chinese family with Congenital heart defects (CHD) caused by a novel *TAB2* nonsense mutation has been published in 2020 (Chen et al., 2020). We additionally tracked the multi-Omics profiles on a chromosomal map. **Figure 3D** shows the chromosomal map for module 13, indicating the combination of two or more different Omics profiles located within a small window on the same chromosome.

These follow-up analyses suggest the biological relevance of the multi-Omics modules resulted from the proposed Cox-sMBPLS algorithm. However, further functional and validation studies (such as *in vivo* validation using animal models) are required to identify novel biomarkers.

DISCUSSION

A survival prediction model (Cox-sMBPLS) based on leveraging and integrating information across multi-Omics compartments via the *cis*-regulatory quantitative effects (eQTL, eQTM, meQTL) was developed. It also enables identification of multi-Omics modules -combinations of different Omics features- exhibiting a large effect on survival probabilities. The proposed modeling framework can easily accommodate a large number of blocks and thus other Omics types with minor modifications.

In the past decade, a large body of literature was developed to introduce methods relating Omics profiles and time to an event such as recurrence in cancer patients, death, etc. Cox-PH (Cox, 1972) is the most widely used method to model the time to such events, for which several high-dimensional adaptations have been proposed in the literature (Park et al., 2002; Tan et al., 2006; Zhang and Zhang, 2020). To also leverage the biological information besides the censoring, we employed the *cis*-regulatory information and a censoring-reweighting technique in our proposed algorithm. The key output of the Cox-sMBPLS is to determine multi-Omics modules that are most associated with disease progression and patient survival.

Simulation studies showed that both the prediction and feature-selection performance of Cox-sMBPLS is significantly better than competing procedures (El-net Cox and RSF) across multiple settings (**Tables 1**, **2**) and in a heart failure study (**Figure 3**). The gene-set enrichment and disease ontology results confirmed biological relevance of the identified multi-Omics modules. Particularly, we found *PDPK1* and *TAB2* associated with HF which have been well documented for playing a crucial role in heart complications (Kuzmanov et al., 2016; Yin et al., 2017; Marrocco et al., 2019; Chen et al., 2020; Cheng et al., 2020).

A direction of future research is to enhance the incorporation of additional prior biological knowledge; e.g., include functional pathway information. On the validation front, analysis of data from animal studies can assist in identifying novel non-coding features prioritized by significant multi-Omics modules.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Data are available from JD upon reasonable request. Requests to access these datasets should be directed to JD, juliod@cop.ufl.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The University of Illinois at Chicago (UIC). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NV and GM designed the study and wrote the manuscript. NV, CM, AD, LC, and JD performed the multi-Omics data preparation and quality control. NV and JD implemented the analysis. All authors read and approved the final manuscript.

FUNDING

The work of GM and NV was supported in part by NIH grant 5U01CA235487-00.

ACKNOWLEDGMENTS

We thank the staff and researchers at the cardiology clinics at UIC, who provided access to the HF multi-Omics data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.701405/full#supplementary-material

Supplementary Section 1 | Supplementary Methods, including Cox-sMBPLS objective function recasting, and model performance measures.

Supplementary Section 2 | Supplementary Results including simulation experiments, additional simulation results, and additional results for the heart failure data.

REFERENCES

Bastien, P., Bertrand, F., Meyer, N., and Maumy-Bertrand, M. (2015). Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics* 31, 397–404. doi: 10.1093/bioinformatics/btu660

Bastien, P., and Tenenhaus, M. (2001). PLS generalised linear regression. application to the analysis of life time data. *Paper Presented at the PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA*, Paris.

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Comput. Stat. Data Anal.* 48, 17–46. doi: 10.1016/j.csda.2004.02.005

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. *J. Stat. Plan. Inference* 143, 1835–1858. doi: 10.1016/j.jspi.2013.05.019

Chambless, L. E., and Diao, G. (2006). Estimation of time−dependent area under the ROC curve for long−term risk prediction. *Stat. Med.* 25, 3474–3486. doi: 10.1002/sim.2299

Chen, J., Yuan, H., Xie, K., Wang, X., Tan, L., Zou, Y., et al. (2020). A novel TAB2 nonsense mutation (p. S149X) causing autosomal dominant congenital heart defects: a case report of a Chinese family. *BMC Cardiovasc. Disord.* 20:27. doi: 10.1186/s12872-019-01322-1

Cheng, A., Neufeld-Kaiser, W., Byers, P. H., and Liu, Y. J. (2020). 6q25. 1 (TAB2) microdeletion is a risk factor for hypoplastic left heart: a c repaseort that expands the phenotype. *BMC Cardiovasc. Disord.* 20:137. doi: 10.1186/s12872-020-01404-5

Chun, H., and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* 72, 3–25. doi: 10.1111/j.1467-9868.2009.00723.x

Chung, D., Chun, H., and Keles, S. (2012). *An Introduction to the 'spls' Package, Version 1.0*. Madison, WI: University of Wisconsin.

Cox, D. R. (1972). Regression models and life−tables. *J. R. Stat. Soc. Series B Methodol.* 34, 187–202.

Datta, S. (2005). Estimating the mean life time using right censored data. *Stat. Methodol.* 2, 65–69. doi: 10.1016/j.stamet.2004.11.003

Duarte, J. D., Kansal, M., Desai, A. A., Riden, K., Arwood, M. J., Yacob, A. A., et al. (2018). Endothelial nitric oxide synthase genotype is associated with pulmonary hypertension severity in left heart failure patients. *Pulm. Circ.* 8:2045894018773049.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* 32, 407–499.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *J. Am. Stat. Assoc.* 89, 122–127.

Gorlov, I., Xiao, X., Mayes, M., Gorlova, O., and Amos, C. (2019). SNP eQTL status and eQTL density in the adjacent region of the SNP are associated with its statistical significance in GWA studies. *BMC Genet.* 20:85. doi: 10.1186/s12863-019-0786-0

Harden, J. J., and Kropko, J. (2019). Simulating duration data for the cox model. *Political Sci. Res. Methods* 7, 921–928. doi: 10.1017/psrm.2018.19

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* 247, 2543–2546. doi: 10.1001/jama.1982.03320430047030

Heagerty, P. J., and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* 61, 92–105. doi: 10.1111/j.0006-341x.2005.030814.x

Hoerl, A. E., Kannard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Commun. Stat. Theor. Methods* 4, 105–123. doi: 10.1080/03610917508548342

Hornung, R., and Wright, M. N. (2019). Block forests: random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics* 20:358. doi: 10.1186/s12859-019-2942-y

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* 2, 841–860.

Jardillier, R., Chatelain, F., and Guyon, L. (2020). Benchmark of lasso-like penalties in the Cox model for TCGA datasets reveal improved performance with pre-filtering and wide differences between cancers. *bioRxiv*[Preprint]. doi: 10.1101/2020.03.09.984070

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Stat.* 12, 531–547. doi: 10.1198/1061860032148

Jones, P. A. (1999). The DNA methylation paradox. *Trends Genet.* 15, 34–37. doi: 10.1016/s0168-9525(98)01636-9

Kass, S. U., Landsberger, N., and Wolffe, A. P. (1997). DNA methylation directs a time-dependent repression of transcription initiation. *Curr. Biol.* 7, 157–165. doi: 10.1016/s0960-9822(97)70086-1

Kuzmanov, U., Guo, H., Buchsbaum, D., Cosme, J., Abbasi, C., Isserlin, R., et al. (2016). Global phosphoproteomic profiling reveals perturbed signaling in a mouse model of dilated cardiomyopathy. *Proc. Natl. Acad. Sci. U. S. A.* 113, 12592–12597. doi: 10.1073/pnas.1606444113

Lee, D., Lee, Y., Pawitan, Y., and Lee, W. (2013). Sparse partial least−squares regression for high−throughput survival data analysis. *Stat. Med.* 32, 5340–5352. doi: 10.1002/sim.5975

Li, W., Zhang, S., Liu, C.-C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* 28, 2458–2466. doi: 10.1093/bioinformatics/bts476

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585.

Mansour, I. N., Bress, A. P., Groo, V., Ismail, S., Wu, G., Patel, S. R., et al. (2016). Circulating procollagen type III N-terminal peptide and mortality risk in African Americans with heart failure. *J. Card. Fail.* 22, 692–699. doi: 10.1016/j.cardfail.2015.12.016

Marrocco, V., Bogomolovas, J., Ehler, E., dos Remedios, C. G., Yu, J., Gao, C., et al. (2019). PKC and PKN in heart disease. *J. Mol. Cell. Cardiol.* 128, 212–226. doi: 10.1016/j.yjmcc.2019.01.029

Min, E. J., and Long, Q. (2020). Sparse multiple co-Inertia analysis with application to integrative analysis of multi-Omics data. *BMC Bioinformatics* 21:141. doi: 10.1186/s12859-020-3455-4

Park, C. S., Cha, H., Kwon, E. J., Jeong, D., Hajjar, R. J., Kranias, E. G., et al. (2012). AAV-mediated knock-down of HRC exacerbates transverse aorta constriction-induced heart failure. *PLoS One* 7:e43282. doi: 10.1371/journal.pone.0043282

Park, P. J., Tian, L., and Kohane, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* 18(suppl_1), S120–S127.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855.

Rahman, M. S., Ambler, G., Choodari-Oskooei, B., and Omar, R. Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Med. Res. Methodol.* 17:60. doi: 10.1186/s12874-017-0336-2

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39:1.

Tan, Y., Shi, L., Hussain, S. M., Xu, J., Tong, W., Frazier, J. M., et al. (2006). Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. *Bioinformatics* 22, 77–87. doi: 10.1093/bioinformatics/bti737

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Methodol.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Uno, H., Cai, T., Tian, L., and Wei, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* 102, 527–537. doi: 10.1198/016214507000000149

Wangen, L., and Kowalski, B. (1989). A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemom.* 3, 3–20. doi: 10.1002/cem.1180030104

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38(suppl_2), W214–W220.

Wold, S., Martens, H., and Wold, H. (1983). "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils*, eds B. Kågström and A. Ruhe (Berlin: Springer), 286–293. doi: 10.1007/bfb0062108

Yin, H., Guo, X., Chen, Y., Steinmetz, R., and Liu, Q. (2017). TAB2 is molecular switch that critically regulates myocardial survival and necroptosis. *Circ. Res.* 121(suppl_1), A468–A468.

Yosefian, I., Mosa Farkhani, E., and Baneshi, M. R. (2015). Application of random forest survival models to increase generalizability of decision trees: a case study in acute myocardial infarction. *Comput. Math. Methods Med.* 2015:576413.

Zhang, W., and Zhang, Y. (2020). Integrated survival analysis of mRNA and microRNA signature of patients with breast cancer based on Cox model. *J. Comput. Biol.* 27, 1486–1494. doi: 10.1089/cmb.2019.0495

Zhernakova, D. V., Deelen, P., Vermaat, M., Van Iterson, M., Van Galen, M., Arindrarto, W., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* 49, 139–145.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.* 15, 265–286.

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

The reviewer HQ declared a shared affiliation with the authors NV, CM, LC, JD, and GM to the handling Editor at time of review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.