mathematics

MDPI

Article

COSMONET: An R Package for Survival Analysis Using Screening-Network Methods

Antonella Iuliano [1,*,†], Annalisa Occhipinti [2,*,†], Claudia Angelini [3,‡], Italia De Feis [3,‡] and Pietro Liò [4,‡]

[1] Dipartimento di Matematica, Informatica ed Economia (DIMIE), Università degli Studi della Basilicata, 85100 Potenza, Italy
[2] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough TS1 3BX, UK
[3] Istituto per le Applicazioni del Calcolo "Mauro Picone" (IAC), Consiglio Nazionale delle Ricerche, 80131 Naples, Italy; c.angelini@iac.cnr.it (C.A.); i.defeis@iac.cnr.it (I.D.F.)
[4] Computer Laboratory, University of Cambridge, Cambridge CB2 1TN, UK; pl219@cam.ac.uk
[*] Correspondence: antonella.iuliano@unibas.it (A.I.); A.Occhipinti@tees.ac.uk (A.O.)
[†] These first authors contributed equally to this work.
[‡] These authors contributed equally to this work.

check for updates

Citation: Iuliano, A.; Occhipinti, A.; Angelini, C.; De Feis, I.; Liò, P. COSMONET: An R Package for Survival Analysis Using Screening-Network Methods. *Mathematics* **2021**, *9*, 3262. https://doi.org/10.3390/math9243262

Academic Editors: Nicoletta Del Buono and Flavia Esposito

Received: 14 November 2021
Accepted: 12 December 2021
Published: 15 December 2021

Abstract: Identifying relevant genomic features that can act as prognostic markers for building predictive survival models is one of the central themes in medical research, affecting the future of personalized medicine and omics technologies. However, the high dimension of genome-wide omic data, the strong correlation among the features, and the low sample size significantly increase the complexity of cancer survival analysis, demanding the development of specific statistical methods and software. Here, we present a novel R package, `COSMONET` (COx Survival Methods based On NETworks), that provides a complete workflow from the pre-processing of omics data to the selection of gene signatures and prediction of survival outcomes. In particular, `COSMONET` implements (i) three different screening approaches to reduce the initial dimension of the data from a high-dimensional space p to a moderate scale d, (ii) a network-penalized Cox regression algorithm to identify the gene signature, (iii) several approaches to determine an optimal cut-off on the prognostic index (PI) to separate high- and low-risk patients, and (iv) a prediction step for patients' risk class based on the evaluation of PIs. Moreover, `COSMONET` provides functions for data pre-processing, visualization, survival prediction, and gene enrichment analysis. We illustrate `COSMONET` through a step-by-step R vignette using two cancer datasets.

Keywords: variable screening; network penalization; survival

1. Introduction

In the last several years, technological advances in the high-throughput quantitative analysis of omics data have provided ample opportunities to investigate the onset and progression mechanisms of several complex diseases, including cancer. International collaborations in large projects, such as The Cancer Genome Atlas (TCGA), which constitutes the core of The Genomic Data Commons (GDC Data Portal) [1,2], the European Genome-Phenome Archive (EGA) [3], and the Gene Expression Omnibus (GEO) [4,5], among many others, have contributed to profiling large tumor sets for different omics layers (genomic, epigenomic, transcriptomic, metabolomics, and proteomic data).

The availability of such a huge volume of cancer omics data has favored the development of novel computational and statistical methods for personalized therapeutic strategies, improving the ability to diagnose, treat, and predict cancer progression. In particular, researchers have devoted significant efforts to developing computational methods to cope with the curse of data dimensionality and variables' correlation, which constitute the two main challenges faced when working with survival and omics data. Such methods are based on penalized regression, statistical boosting, random forest, and, more recently, deep learning, and they have proven their utility in several cancer studies (we refer the reader to [6] for an overview of survival methods using multi-omics data). Additionally, in the last few decades, model prediction data-driven methods have emerged in the context of stochastic processes and particle diffusion models [7–9]. These methods aim to increase the accuracy of machine learning and statistical methods applied to experimental data.

Among the several available approaches, penalization methods constitute a general framework that has the advantage of being easily interpretable, allowing the modeling of several situations, and offering competitive performance. According to [10], we can group penalization methods into two main categories: (i) traditional penalized regression methods and (ii) network-penalized Cox regression methods. In particular, traditional penalized regression methods address the high-dimensionality problem through penalized Cox regression approaches, where the penalty serves to regularize the solution and enforce sparsity. Well-known examples in the context of Cox regression are LASSO penalty [11–13], Elastic-net penalty [14,15], SCAD [16], adaptive Lasso [17], and Dantzig selector [18], among many others. Network-penalized Cox regression methods explicitly incorporate the relationships among the variables in the penalty term. Therefore, they improve the prediction capabilities and better address the inherent structure of omics data. The papers [19,20] introduced the idea using linear models, followed by DrCOX [21,22], AdaLNet [23], Net-Cox [24], $L_{1/2}$ penalty [25], and DegreeCox [26] generalized to Cox regression. In [10], the authors described the advantages and limits of penalized methods in the context of linear models, logistic regression, and Cox regression, and [6] provides a large benchmark study with eleven methods (including penalization approaches) for survival analysis. Moreover, [27–30] offer a perspective for precision medicine and examples of successful applications in cancer studies.

All the approaches presented above can take advantage of and reuse the existing biological information available in databases such as KEGG [31], GO [32], and STRING [33] to guide and improve statistical methods. Such information is often available as networks, with nodes being the genes and edges their connections. In previous papers, we showed how such information can be successfully integrated into Cox regression models; see [34,35]. However, to facilitate the use of widespread network-based methods within the biomedical community, it is necessary to make available software that can handle all the analysis phases, not only the mere implementation of the statistical core.

To this purpose, several R packages have been implemented, including both traditional and network-based penalization approaches for several regression contexts [10]. However, if we limit the search to network-based approaches in the context of Cox regression, the availability of the methods is limited and they are often designed for experts. `glmSparseNet` is an R package implementing the methodology proposed in [36]. It includes network-based regularizers in sparse models when a graph structure can represent the feature space (the gene network is a protein–protein interaction network obtained from STRING). The model uses the network's centrality measures as penalty weights in the regularization process to analyze high-dimensional survival data. The function `NetPredCode.R` [37] provides an implementation of the penalized model described in [38], to obtain accurate prediction models. Specifically, it uses the correlation structure of the feature space (e.g., genes) as a network structure to derive network information included in the penalized model. `NetPredCode.R` implements a three-step approach based on (i) network construction, (ii) cluster approaches to detect modules or pathways, and (iii) development of the final prediction model using the detected modules. In summary, `NetPredCode.R` can apply different combinations of network analyses and penalized regression approaches, but its usage requires prior experience in `R`.

As shown in [39,40], introducing statistical screening in the context of linear regression, generalized linear models, and Cox regression holds great potential in improving the accuracy of the proposed models. Furthermore, variable screening can help to better cope with the high-dimensionality challenges by reducing the dimensionality from a high-

/ultra high-dimensional space to a moderate-dimensional space where the penalization approaches might be more effective. We showed the advantages of merging screening techniques with network-based Cox models in [34,35]. In conclusion, although few tools are already available, there is still a lack of versatile packages that combine screening approaches with penalized regression methods based on network analysis.

Here, we present `COSMONET` (COx Survival Methods based On NETworks), a novel R package inspired by the methodology proposed in [34,35]. The novelty of the statistical model behind `COSMONET` is the combination of screening techniques (i.e., the transformation of data from a high-dimensional space into a low-dimensional space) and network-penalized Cox regression approaches for the selection of significant biomarkers. By combining the biological knowledge related to the disease under investigation and the statistical information derived by the data, the package allows the user to identify new potential biomarkers. The package implements a novel pipeline that selects a subset of genes associated with cancer survival and predicts individual patients' risk by evaluating a prognostic index (PI) on the selected gene signature. In addition, `COSMONET` offers a complete workflow that starts from the pre-processing and normalization steps of omics data and trains the model to identify potential biomarkers, computes PIs, and determines the optimal cut-off to separate high- from low-risk patients. Then, it allows the evaluation of the PIs on a new set of patients (test set) and predicts the patient-specific survival risk. The package also includes advanced visualization options to investigate the data using survival curves, heat-maps, gene pathway networks, Venn diagrams, and correlation plots. Moreover, `COSMONET` has a complete and straightforward step-by-step guide, with a detailed vignette, making the proposed methodology ready to be used within the biomedical community. Overall, it provides to the scientific collectivity a comprehensive set of easy-to-use functions to take full benefit of the increasingly available shared and heterogeneous cancer data. `COSMONET` is publicly available and it can be accessed at http://bioinfo.na.iac.cnr.it/cosmonet/ (accessed on 13 November 2021).

In the following sections, we describe the workflow, the statistical methodology, and the functions available in `COSMONET`. We also illustrate the package's capabilities using two cancer datasets (downloaded from the GEO database and GDC data portal).

2. Materials and Methods

`COSMONET` input data are of the form $\{(\boldsymbol{X}_i, Y_i, \delta_i)\}$, for $i = 1, \ldots, n$. Here, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is the omic profile of the ith patient over p genes. $Y_i = \min(t_i, c_i)$ is the response variable composed of the survival time t_i (i.e., the time until endpoint or last follow-up) and the censoring time c_i, and δ_i is the censoring indicator $I(t_i < c_i)$ (i.e., a 0/1 variable, where 0 indicates that the ith patient was censored at time t_i and 1 that the ith patient had an event at time t_i). Specifically, `COSMONET` takes as input data two numeric matrices $\boldsymbol{X_T} \in R^{n_T \times p}$, the training set T, and $\boldsymbol{X_D} \in R^{n_D \times p}$, the testing set D, and the relative survival information $(\boldsymbol{t}_T, \boldsymbol{\delta}_T)$ and $(\boldsymbol{t}_D, \boldsymbol{\delta}_D)$, respectively. Here, n_T and n_D denotes the number of samples in the two sets of data and p is the number of covariates. The omic data can represent gene expression profiles as measured by microarrays or RNA-seq technologies, or any other numeric genome-wide feature that can be associated with genes or proteins to form numeric matrices. In our examples, we will use gene expression data since they are the most popular choice.

Figure 1 summarizes the steps implemented in `COSMONET`. More precisely, it consists of two phases: (i) the training phase and (ii) the testing phase.

(i) The training phase uses $\boldsymbol{X_T}$ to identify prognostic markers, computes PIs, and defines a cut-off PI^* for defining patient risk classes. The core of the training phase consists of the following steps:

- a **screening approach** to provide an essential dimensional reduction step that allows a transition from the high-dimensionality p to a moderate scale $d < p$ using biological information, data-related information, or combining both pieces of knowledge;

- a **network-penalized Cox regression method** to model observed survival times through genome-wide omic profiles (while accounting for coordinated genes functioning in the form of biological pathways or networks) and identify gene signatures;
- a **procedure** to evaluate the PIs and determine **an optimal cut-off** to separate low- from high-risk patients;
- a **subnetwork analysis** based on gene signatures to visualize new potential genes and biological pathways.

(ii) The testing phase performs survival prediction to evaluate prognostic genes on $\boldsymbol{X_D}$ by using the parameters tuned in the training phase (regression coefficients, gene signatures, and the optimal cut-off for the PIs). This phase uses the log-rank test to compare the Kaplan–Meier curves of the patients in the high- and low-risk groups.

We describe the training and testing phases in detail in the following sections.

Note that, when only a single dataset is available, the function`SplitData()` can perform a random split in the training and testing sets before starting the training phase.

2.1. Training Phase

In the training phase, `COSMONET` performs an (i) optional pre-processing and data normalization step, (ii) variable screening (offering three different approaches plus the possibility to give as the input parameter a list defined by the user), (iii) network construction for the regression analysis (offering two different possibilities), (iv) network-penalized Cox regression analysis, and (v) determination of an optimal cut-off PI^* for the PIs. `COSMONET` includes pathway analysis and heat-map visualizations in the training phase to facilitate the interpretation of the results.

The following sections report the details of each step and Table 1 reports the options available in the package for the screening procedure and the network construction that a user can apply during the process of the training phase.

Table 1. Methodologies and corresponding options implemented in `COSMONET` for the training phase of the model. The table shows the different approaches and algorithms provided in the package for variable screening and network construction.

Methodologies	Options Implemented in `COSMONET`
Screening Approaches	Biomedical-driven screening (BMD) Data-driven screening (DAD) Biomedical-driven + data-driven screening (BMD + DAD) User-specific screening
Network Construction	Functional linkage network (FL) User-specific network

2.1.1. Data Pre-Processing and Normalization

`COSMONET` can perform an optional pre-processing phase (see Figure 1a) consisting of normalization between samples and between different datasets (here denoted *between sample normalization*), when necessary. It does not perform the so-called *within-sample normalization* since such a procedure strongly depends on the type of high-throughput platform used. Therefore, individual samples must be already normalized according to the specific technology.

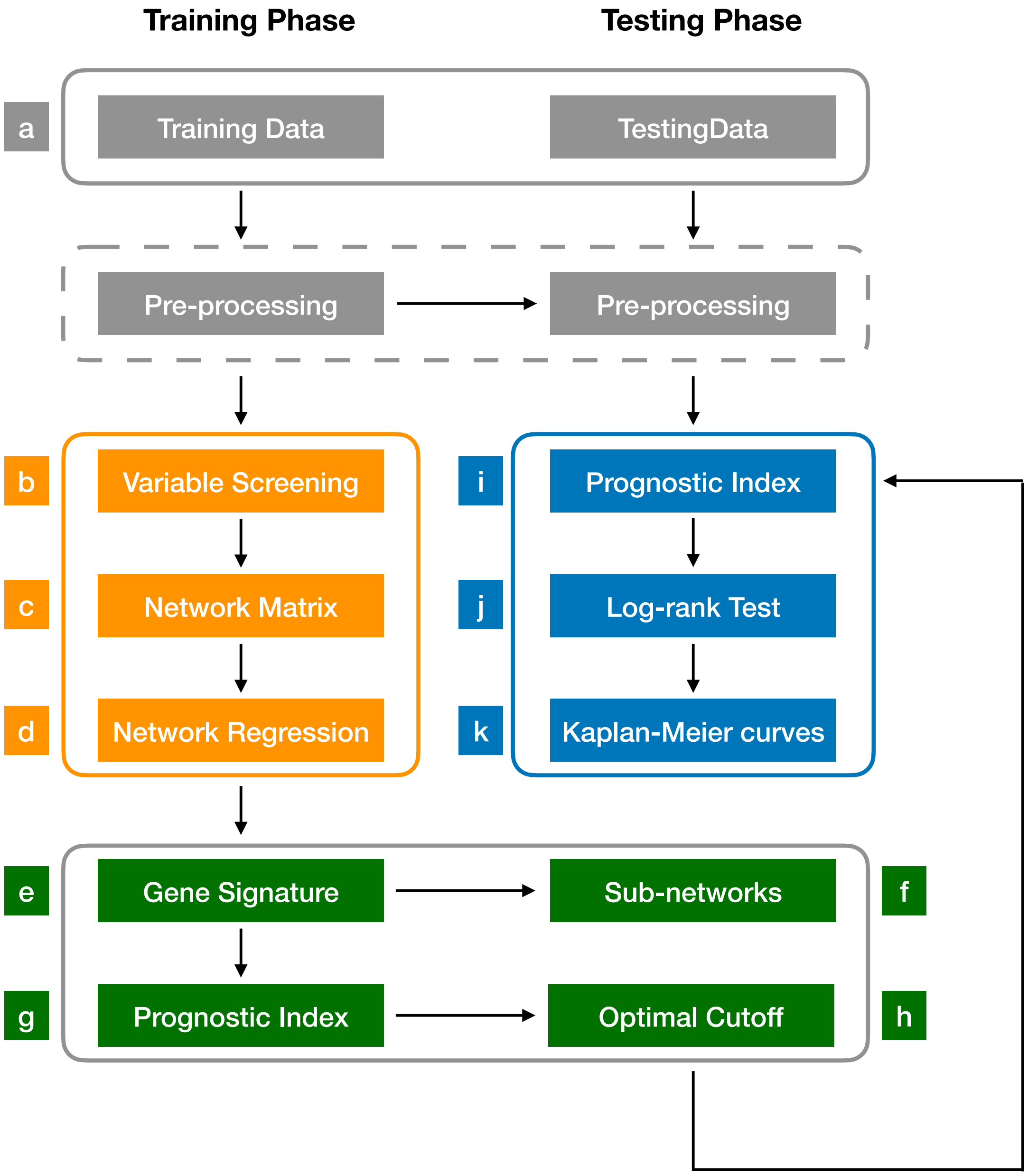

Figure 1. (a) X_T and X_D (with survival information) represent the input data. They can be two distinct datasets or part of a single large dataset. If the dataset is unique, the user can consider the function `SplitData()`, and choose the split percentage, to obtain a random division of the dataset. In both cases, `COSMONET` can perform the normalization between the two sets, if not already normalized. (b) `COSMONET` uses screening approaches on X_T to reduce the dimensionality from a large scale to a moderate scale. (c) `COSMONET` builds the network matrix to incorporate prior biological knowledge into the model. (d) `COSMONET` applies network-based Cox regression methods to select the high-risk cancer genes on the screened subset obtained in step (b). (e) `COSMONET` outputs the gene signatures. (f) `COSMONET` shows pathway sub-networks based on gene signatures. (g) `COSMONET` computes PIs for each patient in X_T. (h) `COSMONET` selects adaptively the optimal cut-off PI_T^* on X_T. (i) `COSMONET` computes PI on X_D using the regression coefficient and the PI_T^* and assigns each patient in the testing set into the high-/low-risk group depending on the comparison between PIs and the estimated threshold PI_T^*, (j) `COSMONET` performs the log-rank test to compare the survival curves between the patients in the high-risk and low-risk groups. (k) `COSMONET` plots Kaplan–Meier curves.

The function `NormalizeBetweenData()` implements the *between-sample normalization*. This function uses *quantile normalization* to make the distributions of the training and testing sets the same across samples. The function first normalizes the training set using the quantile normalization and obtains the normalized training dataset used for the training phase. Then, it normalizes the testing dataset to the normalized training dataset sample-by-sample. It adds one column (i.e., a sample) of the testing set to the training set and normalizes the dataset with $n_T + 1$ samples, repeating the normalization for all n_D samples in the testing set. Then, it takes all the normalized test columns and builds the normalized testing set. The between-sample normalization is necessary to remove unwanted variability, improve the models' performance and stability, and make the two datasets comparable. The proposed approach has the advantage that the normalization training set is entirely independent of the testing set. Moreover, the test set samples are normalized individually to the training, thus allowing the system to incorporate new samples with a study's progress. The user can omit the pre-processing step if the datasets were already normalized using other procedures.

2.1.2. Screening Techniques

The variable *screening approaches* (Figure 1b) reduce the number of variables p to a moderate dimension $d < p$. To this purpose, let $\{x_j, j \in \mathcal{I}\}$ be the subset of the screened variables. Denote $d = |\{x_j, j \in \mathcal{I}\}|$ its cardinality. `COSMONET` includes three different screening approaches.

- **Biomedical-driven (BMD) screening.** The subset $\{x_j, j \in \mathcal{I}\}$ of the screened variables consists of those variables that are known to be associated with the type of cancer under investigation [34]. Typically, this knowledge is derived from the literature or external databases. In [34], we considered the Human Experimental/Functional Mapper (HEFaIMp) database [41]. Recently, the HumanBase database [42]) replaced HEFaIMp. HumanBase uses posterior probabilities (PPs) to identify a significant relationship between a set of genes and the disease of interest. Therefore, to select the *biologically relevant* gene set $\mathcal{I}_{BMD}$, `COSMONET` uses the information provided by HumanBase to (i) decreasingly rank the p genes (connected to the disease of interest) based on PPs and (ii) select the top genes. The user can perform the latter step by using as a threshold score th_{BMD} a cut-off on PPs, i.e., the value of the reduced dimension d selecting the top d genes (the genes with the highest PPs).
- **Data-driven (DAD) screening.** The subset $\{x_j, j \in \mathcal{I}\}$ of the screened variables is obtained by using only information from the data, as in [43]. In particular, it uses component-wise estimators that are computed very efficiently and do not suffer from the numerical instability associated with ultrahigh-dimensional estimation problems, as follows. Let $\mathcal{M}_* = \{1 \leq i \leq d : \beta_i^* \neq 0\}$ be the true sparse Cox model, where $\boldsymbol{\beta}^* = (\beta_0^*, \ldots, \beta_d^*)^T$ denotes the true value of the parameter vector and $\beta_0^* = 0$. The maximum marginal likelihood estimator (MMLE) β_j^M, for $j = 1, \ldots, d$, is defined in Cox models as the maximizer of the log-partial likelihood with a single covariate:

$$\beta_j^M = \arg\max_{\beta_j} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_{ij}^T \boldsymbol{\beta_j} - log \left[\sum_{j \in R(t_i)} exp(\mathbf{x}_{ij}^T \boldsymbol{\beta_j}) \right] \right\}, \quad (1)$$

  where $R(t_i)$ is the risk set.
  This procedure (implemented in the `MarginalCoxRanking()` function) provides the marginal regression coefficients of each feature and the p-values associated with the univariate models. Then, to select the optimal threshold d and optimize data prediction, `COSMONET` allows the user to rank the genes according to the three options below.

  a. Magnitude of the *marginal regression coefficients*. This approach selects the d-top-ranked covariates, i.e., the d genes with the largest marginal coefficients in absolute value. Typically, d is equal to $\lfloor n/logn \rfloor$ [43].
  b. p-value. This option identifies as d-genes all the genes that have p-values < 0.05, regardless of the magnitude of their *marginal regression coefficients*.

c. Magnitude of the *marginal regression coefficients* and p-values. This approach first orders the genes according to the largest marginal regression coefficients in absolute value; then, it selects only those genes that have p-values < 0.05.

- **Biomedical-driven + data-driven (BMD + DAD) screening.** The subset $\{x_j, j \in \mathcal{I}\}$ of the screened variables is obtained by combining the biomedical information and the data-driven knowledge. To this purpose, `COSMONET` takes the union of the BMD and DAD sets of genes. By using BMD + DAD screening, `COSMONET` explores the best model that can sufficiently explain the data in the most parsimonious way to (i) make use of available information, (ii) identify new markers that the BMD screening ignores, and (iii) improve the ability to make precise prognosis, diagnosis, and treatments.

The function `ScreeningMethods()` implements the three methodologies discussed above. Moreover, `COSMONET` also allows the user to import an external list of genes as the screened variables (obtained from the user's clinical experience or the literature).

2.1.3. Gene Network Construction

The set of screened variables $\{x_j, j \in \mathcal{I}\}$ is used to build the networks that will be included in the penalized Cox regression model (Figure 1c,d). In particular, `COSMONET` implements the following network construction procedure:

- **Functional linkage (*FL*) network.** `COSMONET` builds $\mathbf{S} = \left(S_{ij}\right)_{(i,j)\in\mathcal{I}_{BMD}}$ from the HumanBase tool [42,44]). Each element in the S matrix represents the PP that two genes are functionally related. The higher the probability, the stronger the functional relation between the genes in the disease of interest. `COSMONET` completes the S matrix by setting to zero the weights of all genes that are not identified by the functional linkage map. The function `CreateNetworkMatrix()` builds matrix $\mathbf{S}$.

Moreover, `COSMONET` allows the user to import a molecular network $\mathbf{S}$ that matches the set of screened genes $\{x_j, j \in \mathcal{I}\}$. Note that the network matrix $\mathbf{S}$ must be an adjacency matrix with zero diagonal and non-negative off-diagonal.

2.1.4. Network-Penalized Cox Regression Algorithm

The next step consists in the application of a penalized algorithm using the set of screened variables $\{x_j, j \in \mathcal{I}\}$ and the a priori network information (Figure 1b,d).

Given the Cox penalized partial log-likelihood function

$$\ell(\boldsymbol{\beta}_{\mathcal{I}}) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_{\mathcal{I},i}^T \boldsymbol{\beta}_{\mathcal{I}} - log \left[\sum_{j\in R(t_i)} exp(\mathbf{x}_{\mathcal{I},j}^T \boldsymbol{\beta}_{\mathcal{I}}) \right] \right\},$$

`COSMONET` considers the following penalized problem:

$$\min_{\boldsymbol{\beta}_{\mathcal{I}}} \ell(\boldsymbol{\beta}_{\mathcal{I}}) + P_{\rho,\gamma}(\boldsymbol{\beta}_{\mathcal{I}}), \tag{2}$$

with

$$P_{\rho,\gamma}(\boldsymbol{\beta}_{\mathcal{I}}) = \rho \|\boldsymbol{\beta}_{\mathcal{I}}\|_0 + \gamma \Gamma(\boldsymbol{\beta}_{\mathcal{I}}), \tag{3}$$

where $\rho, \gamma > 0$ are two regularization parameters. The penalty function consists of two terms. The first term is a ℓ_0-norm that enforces sparsity in the solution. The second term $\Gamma(\cdot)$ is a Laplacian matrix constraint that gives smoothness among connected genomic variables or regression coefficients in the network. More precisely, the graph Laplacian regularization $\Gamma(\boldsymbol{\beta}_{\mathcal{I}})$ is equal to $\boldsymbol{\beta}_{\mathcal{I}}^T \boldsymbol{L} \boldsymbol{\beta}_{\mathcal{I}}$ with $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S} \in \mathbb{R}^{p\times p}$ (graph Laplacian), where $\boldsymbol{D}$ is the degree matrix and $\boldsymbol{S} \in \mathbb{R}^{p\times p}$ is the adjacency matrix.

To solve the optimization problem proposed in Equation (2), `COSMONET` uses an efficient ℓ_0-norm feature selection algorithm based on augmented and penalized minimization-L_0 (APM-L_0) [45]. Specifically, APM-L_0 employs a two-stage procedure where the first stage replaces the ℓ_0-penalty with a penalty that provides computationally manageable optimization, and the second stage performs hard thresholding. Indeed, the objective function in Equation (2) is reformulated by introducing a surrogate parameter $\boldsymbol{\theta}_{\mathcal{I}}$ of $\boldsymbol{\beta}_{\mathcal{I}}$ and bounding the difference between them by a smooth convex function, which guarantees the convergence of the proximal operator. The following formula gives the Lagrangian form of APM-L_0:

$$L(\boldsymbol{\beta}_{\mathcal{I}}, \boldsymbol{\theta}_{\mathcal{I}}) = \ell(\boldsymbol{\beta}_{\mathcal{I}}) + \rho||\boldsymbol{\theta}_{\mathcal{I}}||_0 + \lambda\left[\alpha\sum_{j=1}^{p}\Phi_j(|\beta_{\mathcal{I}_j} - \theta_{\mathcal{I}_j}|) + (1-\alpha)\Gamma(\boldsymbol{\beta}_{\mathcal{I}})\right], \tag{4}$$

where $\Phi_j(\cdot)$ is a convex function such that $\Phi_j(0) = 0$ and $\Phi_j(|x|) \geq 0$, and $\alpha \in (0,1]$. To minimize Equation (4) for a given λ, ρ and α, APM-L_0 updates all parameters using the following algorithm:

$$\hat{\boldsymbol{\beta}}_{\mathcal{I}} = \text{argmin}_{\boldsymbol{\beta}_{\mathcal{I}}}\ell(\boldsymbol{\beta}_{\mathcal{I}}) + \lambda\left[\alpha\sum_{j=1}^{p}\Phi_j(|\beta_{\mathcal{I}_j} - \theta^0_{\mathcal{I}_j}|) + (1-\alpha)\Gamma(\boldsymbol{\beta}_{\mathcal{I}})\right], \tag{5}$$

$$\hat{\boldsymbol{\theta}}_{\mathcal{I}} = \text{argmin}_{\boldsymbol{\theta}_{\mathcal{I}}}\rho||\boldsymbol{\theta}_{\mathcal{I}}||_0 + \lambda\sum_{j=1}^{p}\Phi_j(|\hat{\beta}_{\mathcal{I}_j} - \theta_{\mathcal{I}_j}|). \tag{6}$$

where $\theta^0_{\mathcal{I}_j}$ is an initial value. Equation (6) is minimized component-wise, and its solution is given by

$$\hat{\theta}_j = \hat{\beta}_j I\Big(\Phi_j(|\hat{\beta}_j|) > \frac{\rho}{\lambda}\Big), \tag{7}$$

i.e., $\hat{\boldsymbol{\theta}}_{\mathcal{I}}$ is obtained by hard thresholding the $\hat{\boldsymbol{\beta}}_{\mathcal{I}}$ parameters obtained in the first step. COSMONET sets $\alpha = 0.5$ as the default parameter, and chooses λ and ρ iteratively by k-fold cross-validation ($k = 5$ as default parameter). Our package allows the user to select either the value of λ that gives the minimum average cross-validated error or the value of λ that gives the most regularized model such that the cross-validated error is within one standard error of the optimal. The latter choice is inspired by the glmnet package [46], and we modified the APM-L_0 package for such a purpose (the modified version is incorporated into COSMONET).

Overall, this step selects a subset of genes (i.e., $\hat{\boldsymbol{\beta}}_{\mathcal{I}} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^T \neq \mathbf{0}_{\mathcal{I}}$, gene signature in Figure 1e), which is used in the next step to determine PIs and predict patients' risk groups.

2.1.5. Prognostic Index (PI) and Survival Analysis

By using the regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq \mathbf{0}_{\mathcal{I}}$, for each patient i ($i = 1, \ldots, n_T$) in the training set T, COSMONET computes PI^T_i, defined as $PI^T_i = \boldsymbol{x}^{\mathcal{I}}_i\hat{\boldsymbol{\beta}}_{\mathcal{I}}$, where $\boldsymbol{x}^{\mathcal{I}}_i$ is the vector of screened gene expression values associated with the i-th patient (Figure 1g). To determine the optimal cut-off $PI^{T,*}$ for dividing the patients into low-risk (LR) and high-risk (HR) groups, COSMONET offers three different procedures (Figure 1h).

- *Adaptive-based approach*: each patient i ($i = 1, ..., n_T$) in T is placed in the high-risk (or low-risk) group if PI^T_i is above (or below) the q_γ-quantile, where

$$q_\gamma = (q_{\gamma_1}, \ldots, q_{\gamma_j}), \text{ for } (\gamma_1, \ldots, \gamma_j) = (0.25, 0.30, \ldots 0.80).$$

  The procedure is repeated in an adaptive manner for each q_γ, and a log-rank test for each q_γ-quantile is used to compare the Kaplan–Meier survival curve between the two risk groups. The optimal cut-off $PI^{T,*}$ is the value that corresponds to the best separation of the two groups, i.e., the q_γ-quantile related to the lowest non-zero p-value resulting from the log-rank test.

- *Median-based approach*: each patient i ($i = 1, \dots, n_T$) in T is set in the high-risk (or low-risk) group if PI_i^T is above (or below) the q_γ-quantile with γ equal to 0.50. The optimal cut-off $PI^{T,*}$ is the value that corresponds to the median of the PIs.
- *Survival-based approach*: each patient in T is allocated to the high-risk (or low-risk) group by using an outcome-oriented method that provides a cut-point corresponding to the most significant relationship with the survival. `COSMONET` uses the `surv_cutpoint()` function from the `survminer` package.

The function `SelectOptimalCutoff()` computes the optimal cut-off $PI^{T,*}$ and the p-value corresponding to the Kaplan–Meier (KM) survival curves on T. The KM curves provide a useful visual representation for assessing the effect of cancer progression over time for different groups of patients.

The function `CosmonetTraining()` performs the complete procedure (variable selection and survival analysis on T).

2.1.6. Network and Pathway Analysis

`COSMONET` provides an interactive visualization to investigate the gene signature $\hat{\beta}_{\mathcal{I}} \neq \mathbf{0}_{\mathcal{I}}$ in terms of pathway analysis (Figure 1f). It uses pathway information from the KEGG database [31] to create sub-networks of the gene signature identified by the regression algorithm (Figure 1e). Specifically, `COSMONET` generates a dashboard with two networks: (i) a sub-network of the *not-isolated* genes, i.e., genes that share at least one pathway with another selected gene, and (ii) a complete network displaying the full gene signature selected by the network regression method. Each vertex in the network is a gene. An edge between two genes indicates that they belong to the same KEGG pathway. The color of the nodes/genes allows identification of the functional link between the genes and the disease of interest according to the HumanBase database. Specifically, `COSMONET` automatically retrieves from HumanBase a list of genes ranked according to the PP of being associated with the disease under investigation (identified through the disease ID according to the Disease Ontology (DO) Project [47]). The signature genes among the top 500 genes in the list are displayed in red (and labeled as Mapped Up); the other genes in the list are shown in blue (and labeled as Mapped Down). Any gene signature not recorded in the HumanBase ranking is shown in white (and labeled as Not Mapped).

The complete networks can be useful to gauge an understanding of the type of genes selected as gene signatures by the different approaches (mapped up, mapped down, or not mapped genes). Specifically, the networks allow the detection of any not mapped isolated genes (white nodes, i.e., genes that are not associated with cancer under study by HumanBase), enabling the identification of new potential biomarkers.

The function `GenerateNetwork()` implements the pipeline for visualizing the gene pathway network. This function can also be used as a standalone tool for visualizing the network given a list of genes provided by the user. `COSMONET` includes information from the KEGG database in an internal file (KEGGrepository.RData) that is used by the `GenerateNetwork()` function to automatically map the list of genes to the corresponding KEGG pathways, and create a gene-by-pathway adjacency matrix. This matrix is then used to create the final gene pathway network.

2.1.7. Heat-Maps

To facilitate the interpretation of the results, `COSMONET` includes the `heat-mapSurvGroup()` function. This function first orders patients according to PIs (from the highest to the smallest). Then, it divides them into two risk groups (i.e., low-risk and high-risk classes) using the optimal cut-off PI^*. Finally, it depicts the expression values for each gene of the signature as a Z-score. In this way, the function provides a practical way to identify genes whose upregulation leads to a poorer prognosis (i.e., those in red in the high-risk group) and those whose downregulation leads to a poorer prognosis (i.e., those depicted in green in the high-risk group). Therefore, the heat-map in clinical studies aids in visualizing and pointing out the gene signature (and up/down regulation of key genes), assessing its reproducibility between the training and testing sets (and other datasets), and evaluating the similarity between observations or clusters of patients, in a single figure.

2.2. Testing Phase

The testing phase works on an external dataset, or the testing set resulting from the initial splitting into training and testing sets, as explained in Section 2.1.1. In the testing phase, survival analysis is performed to assess the prediction accuracy (Figure 1i–k).

Survival Analysis

To perform survival analysis, `COSMONET` computes PIs for each patient j ($j = 1, \ldots, n_D$) in the test set as $PI_j^D = \boldsymbol{x}_j^{\mathcal{I}} \hat{\boldsymbol{\beta}}_{\mathcal{I}}$, where $\boldsymbol{x}_j^{\mathcal{I}}$ is the vector of screened gene expression values associated with the j-th patient and $\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq \mathbf{0}_{\mathcal{I}}$ is the gene signature derived from the training phase (Figure 1i). The optimal cut-off $PI^{T,*}$ selected on the training set T is used on the validation set to split the patients into high-risk and low-risk groups. Each j-th patient is designated as a high- (or low-) risk of death if PI_j^D is above (or below) the optimal cut-off $PI^{T,*}$. `COSMONET` uses the log-rank test to calculate the statistical significance level (i.e., computing p-values; Figure 1j) and the KM curves to further analyze the results (Figure 1k). A good separation between high- and low-risk survival curves and a significant p-value (i.e., p-value < 0.05) indicate that the prognostic classifiers selected in the validation set can identify good prognosis patients. In other words, the low-risk class has a significantly better prognosis than the high-risk class. The function `ValidationTest()` implements the procedure to perform survival analysis using the testing set. The function `CosmonetTesting()` performs the complete prediction procedure. The KM curves allow us to visualize the survival probability depending on the assigned risk class and evaluate the difference. The larger is the difference between the two curves, the better is the gene signature as a prognostic biomarker.

2.3. Correlation Analysis

The function `CorrPlot()` allows the user to analyze and visualize the correlation coefficient R (and its p-value) between results obtained using two pairs of PIs, such as those generated using different screening techniques on the same dataset. This enables the comparison of risk assignments based on the two gene signatures and allows us to classify the patients at the individual level: low-risk, high-risk, and not consistently identified (usually, they are borderline). From a clinical point of view, we suggest that the group of patients not consistently recognized might need further investigation to establish cancer risk factors. The function allows the user to select different methods for computing correlation coefficients, such as Pearson, Kendall, or Spearman. The benefit of using correlation analysis is to assess the overall concordance between the PI indexes obtained using different methods, and to easily identify patients that are consistently assigned to one class or the other, from patients whose assignment might depend on the chosen method. In the latter case, the final assignment might be done using additional information.

2.4. `R` *Implementation*

`COSMONET` is a novel `R`-package that implements all the steps described in Figure 1. It is available at http://bioinfo.na.iac.cnr.it/cosmonet/ (accessed on 13 November 2021), as a source code, with data examples and a detailed vignette. The main function `Cosmonet()` can be used to run the full pipeline. However, the user can also run the two phases individually by using the `CosmonetTraining()` and `CosmonetTesting()` functions. Table 2 describes all the functions available in the package.

`COSMONET` uses the APM-L_0 procedure implemented in the `APML0 package` (https://cran.r-project.org/web/packages/APML0/index.html, accessed on 13 November 2021) to fit the penalized regression, and the `Survminer` package to compute the Kaplan–Meier survival curves and estimate the log-rank test p-value. The pathway analysis and network visualization use the KEGG database, the `igraph` (https://cran.r-project.org/web/pac kages/igraph/index.html, accessed on 13 November 2021), and the `visNetwork` packages (https://cran.r-project.org/web/packages/visNetwork/index.html, accessed on 13 November 2021).

Table 2. Summary of the `COSMONET` R package functions based on Figure 1.

Step	Function	Description

a	`SplitData()`
	`NormalizeBetweenData()`
		Split data into training T and testing D set.
		Perform normalization between training T and testing D set.

b	`MarginalCoxRanking()`
	`ScreeningMethod()`
		Apply ranking by marginal utility.
		Perform screening methods, i.e., BMD, DAD, and BMD + DAD screening.

c	`CreateNetworkMatrix()`	Create adjacency matrix with zero diagonal and non-negative off-diagonal matrix using cancer-specific genes.

d–h	`CosmonetTraining()`
	`SelectOptimalCutoff()`
	`GenerateNetwork()`
	`HeatmapSurvGroup()`
		Fit network-regularized Cox regression model to identify the signature genes $\hat{\beta}_{\mathcal{I}} \neq 0$ on T.
		Select optimal cut-off $PI^{T,*}$ on training set T; generate Kaplan–Meier curves resulting from the log-rank test, distribution plot of PI^{T}.
		Generate a biological network based on KEGG pathways to investigate the signature genes.
		Plot hierarchical clustering heat-map of signature genes of low- and high-risk prognostic survival groups using training T and testing D set.

i–k	`CosmonetTesting()`
	`ValidationTest()`
		Make prediction on data.
		Return test validation, Kaplan–Meier curves resulting from the log-rank test on D using the signature genes $\hat{\beta}_{\mathcal{I}} \neq 0$ and the optimal cut-off $PI^{T,*}$, distribution plot of PI^{T}.

d–k	`Cosmonet()`	Used to run the full pipeline including both `CosmonetTraining()` and `CosmonetTesting()`.

Utility function	`CorrPlot()`
	`VennPlot()`
		Create a scatter plot including the correlation coefficient, p-value and linear regression line between the prognostic indices PIs.
		Display three-set Venn diagram plots between the number of patients at low-risk (or high-risk) obtained for each screening.

2.5. Real Data Examples

To illustrate the features and performance of `COSMONET`, we used two gene expression data examples in breast and lung cancer studies. Table 3 shows the details of the datasets.

Table 3. Summary of the datasets used in the examples.

Accession Number	Platform	Genes	Samples	Survival Data
GSE2034	Affymetrix HG-U133A	13229	286	Relapse-Free Survival (RFS)
GSE2990	Affymetrix HG-U133A	13229	187	Relapse-Free Survival (RFS)
TCGA-LUAD	Illumina HiSeq	19988	492	Overall Survival (OS)

2.5.1. Example 1: Breast Cancer Datasets from Microarray Case Studies

We used two independent gene expression datasets on breast cancer available from the GEO database [4]. We used GSE2034 as the training set T to build the model, while we used GSE2990 as the testing set D to validate the model. The training set T consists of gene expression profiles from the total frozen RNA of 286 lymph-node-negative breast cancer patients [48]. The testing set D contains gene expression profiles of 189 invasive breast carcinomas [49]. The median survival time (RFS) in T was 86 months, and the censoring proportion was 62.59%, while the median survival time (RFS) in D was 77 months and the censoring proportion was 63.49%.

We used the `RMA` and `preprocessCore` Bioconductor packages for the pre-processing steps (see Section 2.1.1). After the within-arrays normalization and the reduction to the same p-dimensional feature space (for a total of 13229 genes), we applied the function `NormalizeBetweenData()` to perform the normalization between datasets (see Supplementary Figure S1A).

2.5.2. Example 2: Lung Cancer Dataset from an RNA-Seq Case Study

We considered TCGA-LUAD (lung adenocarcinoma) gene expression data from the GDC Data Portal [1]. The data were obtained from the Illumina HiSeq platform and included 492 lung cancer patients (for a total of 19988 genes). We used the data already pre-processed and normalized (gene-level, RPKM), which are available in the LinkedOmics portal [50]. The median survival time (OS) in T was 701 days, and the censoring proportion was 63.82%, while the median survival time (OS) in D was 624.5 days, and the censoring proportion was 63.82%.

To analyze the dataset, we first randomly split the dataset into 50% train (246 samples) and 50% test (246 samples) and then applied the between normalization by using the functions `splitData()` and `NormalizeBetweenData()`, respectively (see Supplementary Figure S1B).

3. Results

This section aims to illustrate the functionalities of `COSMONET` using microarray and RNA-Seq data examples. The accompanying on-line vignette, available at http://bioinfo.na.iac.cnr.it/cosmonet/, provides a step-by-step analysis of the following examples (accessed on 14 November 2021). Users can also access the pre-processed data and the repositories containing the prior biological information. `COSMONET` is based on the Disease Ontology (DO) Project [47] to identify the disease ID for the disease of interest (e.g., breast cancer DOID:1612 and lung cancer DOID:1324). The normalization between the training set T and the testing set D (Table 2, step (a)) was performed as described in Sections 2.5.1 and 2.5.2, respectively (see Supplementary Figure S1).

3.1. Breast Cancer Microarray Example Results

3.1.1. Results Using the BMD Screening

We applied the BMD screening procedure by intersecting the 500 genes with the highest PPs associated with breast cancer (according to HumanBase) with the 13229 genes measured on the gene expression microarrays (Table 2, step (b)). As a result, the screening retained 437 BMD genes (i.e., $|\mathcal{I}_{BMD}| = 437$). In the training phase, `COSMONET` identified

40 BMD genes as potential breast cancer signature genes (with $\alpha = 0.5$), which we used to compute $PI^{T,i}$ $i = 1, \ldots, n_T$ and the optimal cut-off $PI^{T,*}_{BMD}$ (see Table S1). Here, for illustrative purposes, we used the median-based procedure (q_γ-quantile with $\gamma = 0.50$), i.e., $n_{T_{LR}} = n_{T_{HR}} = 143$ (Table 2, steps (d)–(h)). The on-line vignette provides additional results and plots on the training set T.

In the testing phase, the estimation of $PI^{D,i}$, $i = 1, \ldots, n_D$ led to $n_{D_{LR}} = 144$ low-risk patients and $n_{D_{HR}} = 43$ high-risk patients (Table 2, steps (i)–(k)). The on-line vignette provides additional results and plots on the testing set D. Table 4 summarizes the results.

Table 4. Breast cancer data. Summary of BMD, DAD, and BMD + DAD screening network Cox regression methods. The table is divided into two sides corresponding to the training and testing phases. In the training phase are indicated the type of screening, the number of screened genes $\mathcal{I}$, the signature genes ($\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$), and the number of patients at low and high risk (LR and HR). In the testing phase, we report the p-values for the log-rank test and the number of patients in the high- and low-risk groups.

	Training Phase				Testing Phase		
Screening	Screened Genes	Signature Genes	Low-Risk (LR) Group	High-Risk (HR) Group	p-Value	Low-Risk (LR) Group	High-Risk (HR) Group
BMD	437	40	143	143	0.000156	144	43
DAD	50	32	177	109	5.782×10^{-5}	113	74
BMD + DAD	549	67	143	143	0.0001347	97	90

3.1.2. Results Using the DAD Screening

We also applied the DAD screening by ordering the marginal Cox regression coefficients in absolute value and choosing the $d = 50$ top genes (i.e., $|\mathcal{I}_{DAD}| = 50$) as screened DAD genes (Table 2, step (b)). Note that such a choice is suggested by d equal to $\lfloor n/logn \rfloor$, with $n = 286$. `COSMONET` identified 32 DAD genes (i.e., with $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{DAD}} \neq 0$) as possible breast cancer signature genes (see Table S1). Similarly to the BMD case, we computed $PI^{T,i}$ and the optimal cut-off $PI^{T,*}_{DAD}$. In this case, for expository purposes, we used the survival-based approach. The number of patients at low risk was $n_{T_{LR}} = 177$, while the number of patients at high risk was $n_{T_{HR}} = 109$ (Table 2, steps (d)–(h)). The on-line vignette provides additional results and plots on the training set T.

In the testing phase, $PI^{D,i}$, $i = 1, \ldots, n_D$, we separated the patients into two prognostic groups of cardinality, $n_{D_{LR}} = 113$ and $n_{D_{HR}} = 74$ (Table 2, steps (i)–(k)). The on-line vignette provides additional results and plots on the testing set D. Table 4 summarizes the results.

3.1.3. Results Using the BMD + DAD Screening

The BMD + DAD screening started with 549 screened genes, i.e., $|\mathcal{I}_{BMD+DAD}| = 549$, obtained from the union of the 437 genes identified by the BMD screening and the 50 genes selected by the DAD screening (Table 2, step (b)). From these genes, `COSMONET` identified 67 BMD + DAD genes ($\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD+DAD}} \neq 0$) as potential breast cancer signature genes (see Table S1). We used this signature to compute $PI^{T,i}$ $i = 1, \ldots, n_T$, and the optimal cut-off $PI^{T,*}_{BMD+DAD}$ using the median-based approach. The number of patients at low and high risk was $n_{T_{LR}} = n_{T_{HR}} = 143$ (Table 2, steps (d)–(h)). Survival curves of the training set T and a distribution plot of $PI^{T}_{BMD+DAD}$ are shown in Supplementary Figure S2 (panel A and B as illustrative examples).

In the testing phase, the $PI^{D,i}$ divided the patients into two sets with $n_{D_{LR}} = 97$ and $n_{D_{HR}} = 90$ patients (Table 2, steps (i)–(k)). For illustrative purposes, Figure 2 (panel A) shows the survival curves and the statistically significant p-value (p-value < 0.05) computed using the $BMD + DAD$ genes on the testing set D. Figure 2 (panel B) shows the distribution plot of the PI^{D}_{BM+DAD}. Table 4 summarizes the results.

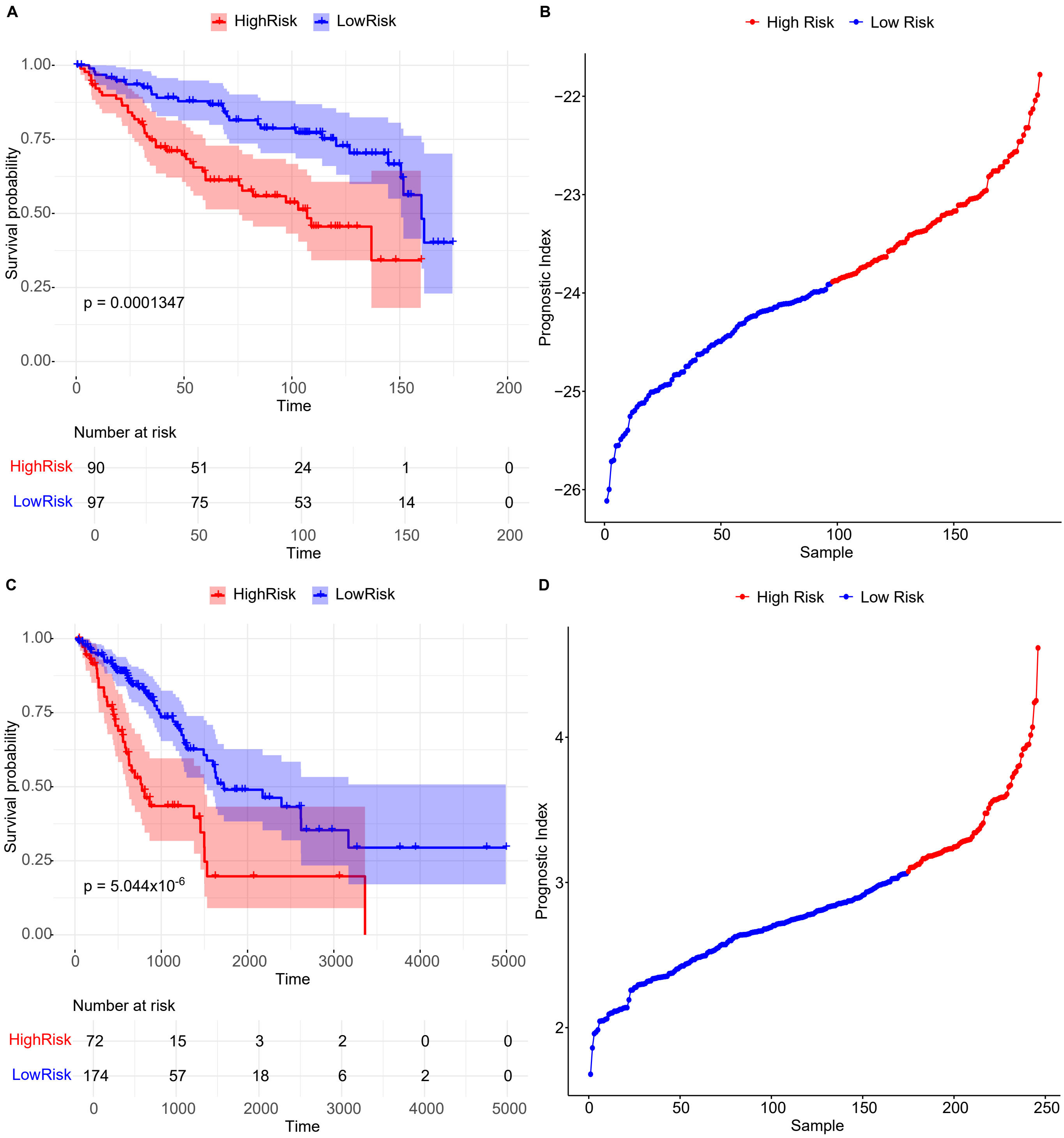

Figure 2. Testing phase. (**A**) Kaplan–Meier curves and (**B**) distribution of the risk score PI^D using the BMD + DAD gene signatures detected in breast cancer. (**C**) Kaplan–Meier curves and (**D**) distribution of the risk score PI^D using the BMD + DAD gene signatures selected in lung cancer. The p-value is statistically significant (p-value < 0.05). The red line (high-risk, i.e., poor survival) and the blue line (low-risk, i.e., high survival) indicate the two risk conditions associated with cancer survival patients in the testing set D. The risk table shows the number of individuals at risk during the time of follow-up.

3.1.4. Interpreting Gene Signatures

To better understand and investigate the gene signatures, i.e., $\hat{\beta}_{\mathcal{I}} \neq 0$, obtained by applying one of the screening network methods, we used the heat.map function (i.e., the function `heat-mapSurvGroups()` in Table 2, steps (d)–(h)). Figure S3 (panel A) reports the heat-maps generated using the BMD + DAD gene signature ($\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD}} \neq 0$).

3.1.5. Correlation Results

We performed a correlation analysis (based on Spearman rank correlation) on PIs obtained using the different screening approaches (PI_{BMD}, PI_{DAD} and $PI_{BMD+DAD}$) on the training T and testing D set. As an illustrative example, Figure 3 (panel A) shows a moderate positive correlation $R = 0.49$ between PI^{D}_{BMD} and PI^{D}_{DAD}, and a strong correlation $R = 0.70$ between PI^{D}_{BMD} and $PI^{D}_{BMD+DAD}$. Both correlations are strongly statistically significant (p-value < 0.05). Points in Figure 3 depicted in red or in blue are consistently assigned to the same risk group, regardless of the screening method; those in grey (in the border of the two classes) suggested that further information might be required before making a clinical decision.

3.1.6. Pathway Analysis

Figure 4 reports the KEGG pathway networks of the not-isolated genes selected by the three screening methods. It is worth noting that the BMD + DAD screening approach (Figure 4 (panel A)) can detect pathway-based relationships between genes that are already known to be biologically associated with the disease (red genes, mapped up) and genes that are not known to be highly associated with breast cancer (blue genes, mapped down). For example, the link between the mapped up gene *CDKN2A* and the mapped down gene *DCC* illustrates that both genes can play an important role in breast cancer as part of the same KEGG pathway (i.e., KEGG pathways in cancer). In fact, several studies have shown the relevance of *CDKN2A* in breast cancer, but only recently, *DCC* has been associated with an increased risk of various cancers [9].

Figure 4 (panel A and B) show that the BMD + DAD and BMD models led to the selection of similar KEGG pathways (they share 13 pathways). Among them, the cytokine–cytokine receptor interaction pathway was identified by the three models. Indeed, cytokine receptors are deeply involved in cancer regulation stem cells through complex interactions with the tumor microenvironment and represent attractive targets for therapeutic development [51,52]. Both the BMD + DAD and BMD models (Figure 4 (panel A and B), respectively) identified three critical genes in breast cancer, i.e., *VEGFC*, *PGF*, and *GAB1*. However, the mapped down genes identified by the BMD + DAD model (Figure 4 (panel A)) allow the detection of potential new gene signatures and cancer pathways that are still unexplored. The interactive networks dashboard is available in Appendix A.1).

3.2. *Lung Cancer, RNA-Seq Example Results*

3.2.1. Results Using the BMD Screening

In this analysis, we selected the top 500 genes associated with lung cancer, downloaded from the HumanBase database (the disease ID is DOID:1324). The BMD screening selected 482 genes (i.e., $|\mathcal{I}_{BMD}| = 482$) as the intersection of the 500 genes with the 19988 genes measured using the RNA-seq technology, i.e., Illumina HiSeq (Table 2, step (b)). Then, applying the penalized network regression, `COSMONET` identified 25 BMD genes (i.e., with $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD}} \neq 0$ and $\alpha = 0.1$) as potential lung cancer signatures (see Table S2). We used these genes to compute $PI^{T,i}_{BMD}$ with $i = 1, \ldots, n_T$ and select the optimal cut-off. Using the median-based approach, we obtained $n_{T_{LR}} = n_{T_{HR}} = 123$ (Table 2, steps (d)–(h)). The on-line vignette provides additional results and plots on the training set T.

In the testing phase, we obtained two prognostic groups, $n_{D_{LR}} = 125$ and $n_{D_{HR}} = 121$ (Table 2, steps (i)–(k)). The on-line vignette provides additional results and plots on the testing set D. Table 5 summarizes the results.

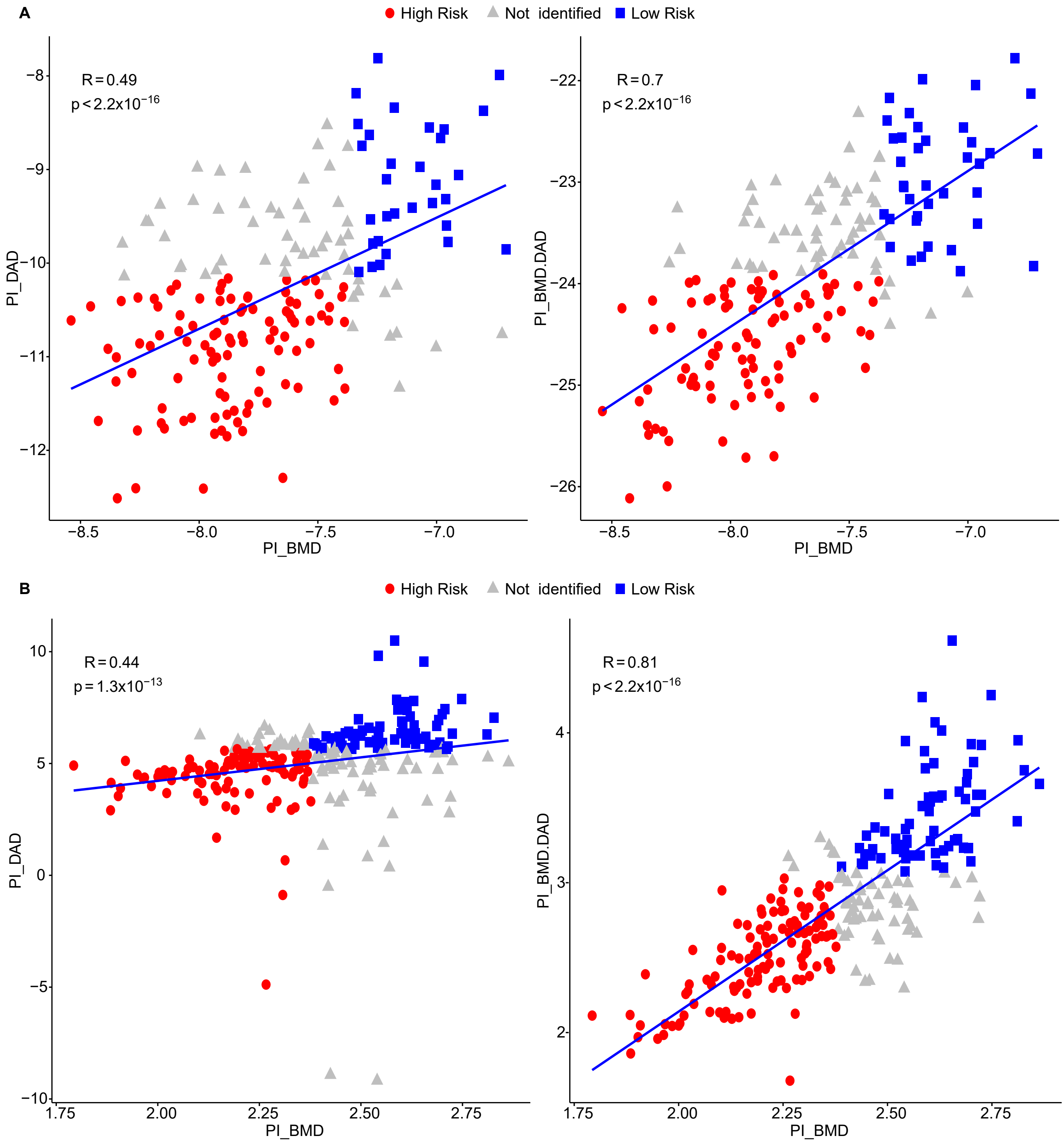

Figure 3. PI correlation plots on testing set. (**A**) Breast cancer. (**B**) Lung cancer. All plots report the correlation coefficient R and the p-value. Red color indicates the high-risk group, blue color indicates the low-risk group, and grey color indicates the not-identified group according to the optimal cut-off $PI^{T,*}$ in each set.

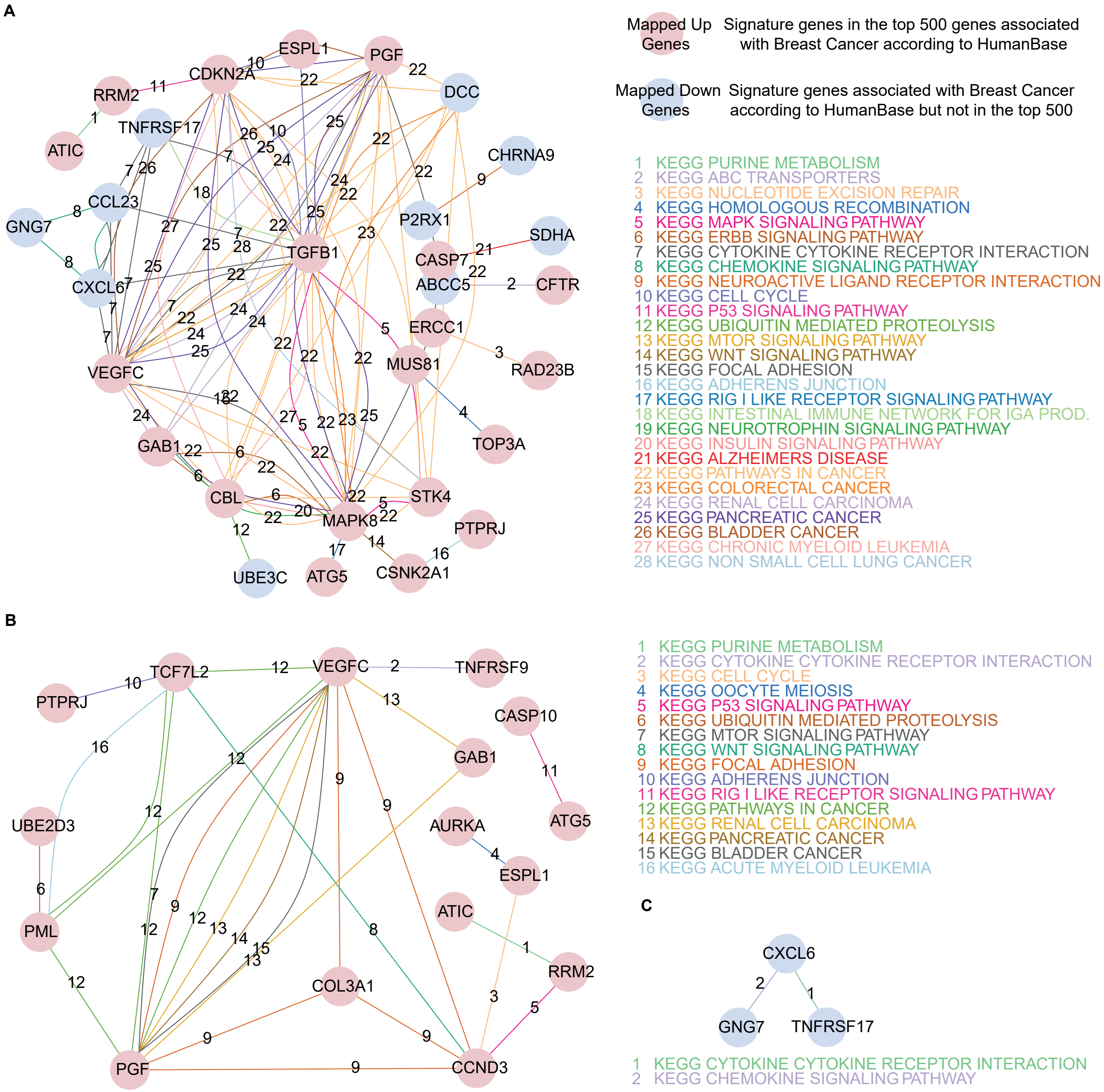
A
Mapped Up Genes
Signature genes in the top 500 genes associated with Breast Cancer according to HumanBase
Mapped Down Genes
Signature genes associated with Breast Cancer according to HumanBase but not in the top 500
ESPL1
PGF
CDKN2A
RRM2
DCC
ATIC
TNFRSF17
CHRNA9
CCL23
GNG7
P2RX1
SDHA
TGFB1
CASP7
CXCL6
ABCC5
CFTR
ERCC1
MUS81
RAD23B
VEGFC
TOP3A
GAB1
STK4
CBL
MAPK8
PTPRJ
UBE3C
ATG5
CSNK2A1
1 KEGG PURINE METABOLISM
2 KEGG ABC TRANSPORTERS
3 KEGG NUCLEOTIDE EXCISION REPAIR
4 KEGG HOMOLOGOUS RECOMBINATION
5 KEGG MAPK SIGNALING PATHWAY
6 KEGG ERBB SIGNALING PATHWAY
7 KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
8 KEGG CHEMOKINE SIGNALING PATHWAY
9 KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION
10 KEGG CELL CYCLE
11 KEGG P53 SIGNALING PATHWAY
12 KEGG UBIQUITIN MEDIATED PROTEOLYSIS
13 KEGG MTOR SIGNALING PATHWAY
14 KEGG WNT SIGNALING PATHWAY
15 KEGG FOCAL ADHESION
16 KEGG ADHERENS JUNCTION
17 KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY
18 KEGG INTESTINAL IMMUNE NETWORK FOR IGA PROD.
19 KEGG NEUROTROPHIN SIGNALING PATHWAY
20 KEGG INSULIN SIGNALING PATHWAY
21 KEGG ALZHEIMERS DISEASE
22 KEGG PATHWAYS IN CANCER
23 KEGG COLORECTAL CANCER
24 KEGG RENAL CELL CARCINOMA
25 KEGG PANCREATIC CANCER
26 KEGG BLADDER CANCER
27 KEGG CHRONIC MYELOID LEUKEMIA
28 KEGG NON SMALL CELL LUNG CANCER
B
TCF7L2
VEGFC
TNFRSF9
PTPRJ
CASP10
GAB1
UBE2D3
AURKA
ATG5
ESPL1
PML
ATIC
RRM2
COL3A1
PGF
CCND3
1 KEGG PURINE METABOLISM
2 KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
3 KEGG CELL CYCLE
4 KEGG OOCYTE MEIOSIS
5 KEGG P53 SIGNALING PATHWAY
6 KEGG UBIQUITIN MEDIATED PROTEOLYSIS
7 KEGG MTOR SIGNALING PATHWAY
8 KEGG WNT SIGNALING PATHWAY
9 KEGG FOCAL ADHESION
10 KEGG ADHERENS JUNCTION
11 KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY
12 KEGG PATHWAYS IN CANCER
13 KEGG RENAL CELL CARCINOMA
14 KEGG PANCREATIC CANCER
15 KEGG BLADDER CANCER
16 KEGG ACUTE MYELOID LEUKEMIA
C
CXCL6
GNG7
TNFRSF17
1 KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
2 KEGG CHEMOKINE SIGNALING PATHWAY

Figure 4. Breast cancer KEGG pathway analysis. KEGG pathway networks of the not-isolated gene signatures selected by BMD + DAD (**A**), BMD (**B**), and DAD (**C**) screening approaches. Red nodes represent the mapped up genes, i.e., signatures that are among the top 500 genes biologically known to be associated with breast cancer according to HumanBase. Blue nodes represent the mapped down genes, i.e., signatures that are biologically known to be associated with breast cancer, but not among the top 500 genes according to HumanBase. Any gene that is not associated with breast cancer will be displayed as a white node (none were selected in the not-isolated gene signatures shown in the figure). The BMD + DAD network (**A**) shows that merging biological and data-driven information allows the identification of pathway-based connections of breast cancer biomarkers (such as TGFB1 and MAPK8) that are not identified by the BMD or DAD approaches, (**B**,**C**), respectively. COSMONET generates the KEGG pathway networks as part of a fully interactive dashboard, which also contains the full networks comprising all the gene signatures, including the isolated genes. The full dashboards are available in the Supplementary Materials.

Table 5. Lung cancer data. Summary of BMD, DAD, and BMD + DAD screening network Cox regression methods. The table is divided into the training and testing phase. In the training phase are indicated the type of screening, the number of screened genes $\mathcal{I}$, the signature genes ($\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$), and the number of patients at low and high risk (LR and HR). In the testing phase are reported the resulting p-values for the log-rank test and the number of patients in the high- and low-risk groups.

Training Phase					Testing Phase		
Screening	Screened Genes	Signature Genes	Low-Risk (LR) Group	High-Risk (HR) Group	p-Value	Low-Risk (LR) Group	High-Risk (HR) Group
BMD	482	25	123	123	0.0007573	125	121
DAD	70	59	151	95	0.003284	154	92
BMD + DAD	567	54	172	74	5.044×10^{-6}	172	74

3.2.2. Results Using the DAD Screening

Here, we performed the DAD screening by ordering the marginal Cox regression coefficients in absolute value and selecting the d top 70 ranked genes (i.e., $|\mathcal{I}_{DAD}| = 70$) as screened DAD genes (Table 2, step (b)). Then, we applied the network-penalized regression, which selected 59 DAD genes (i.e., with $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{DAD}} \neq 0$ and $\alpha = 0.1$) as potential lung cancer signature genes (see Table S2). The DAD signature genes allowed us to compute PI_s and the optimal cut-off. Here, using the survival-based approach, we obtained that the number of patients was $n_{T_{LR}} = 151$ for the low-risk group and $n_{T_{HR}} = 95$ for the high-risk group (Table 2, steps (d)–(h)). The on-line vignette provides additional results and plots on the training set T.

In the testing phase, `COSMONET` separated the patients into two prognostic groups, $n_{D_{LR}} = 154$ and $n_{D_{HR}} = 92$ (Table 2, steps (i)–(k)). The on-line vignette provides additional results and plots on the testing set D. Table 5 summarizes the results.

3.2.3. Results Using the BMD + DAD Screening

As previously done with the microarray data, we first selected as BMD + DAD-screened genes the union of the 482 BMD-screened genes and the 70 ordered DAD-screened genes for a total of 567 genes ($|\mathcal{I}_{BMD+DAD}| = 567$). Then, from the training phase, as before, we detected 54 BMD + DAD genes, i.e., with $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD+DAD}} \neq 0$ and $\alpha = 0.1$ (see Table S2). Here, the number of patients at low and high risk according to the adaptive-based approach was equal to $n_{D_{LR}} = 172$ and $n_{D_{HR}} = 74$, respectively (Table 2, steps (d)–(h)). See Supplementary Figure S2, panel C (Kaplan–Meier curves) and D (distribution plot of $PI^{T}_{BMD+DAD}$).

We validated the survival model in the testing phase, obtaining a subdivision of the patients into two sets of cardinality, $n_{D_{LR}} = 172$ and $n_{D_{HR}} = 74$ (Table 2, steps (i)–(k)). For illustrative purposes, Figure 3 (panel C) shows the Kaplan–Meier curves for OS patients in the low-risk versus high-risk groups together with the p-values (p-values < 0.05). We show the distribution plot of $PI^{D}_{BMD+DAD}$ in Figure 3 (panel D). Table 5 summarizes all the results.

3.2.4. Correlation Results

Furthermore, in this case, we compared the estimated PIs (Spearman rank correlation) on the training T and testing D set obtained using the different screening procedures. For illustrative purposes, Figure 3 (panel B) shows a moderate positive correlation between PI_{BMD} and PI_{DAD} ($R = 0.44$) and a strongly positive correlation between PI_{BMD} and $PI_{BMD+DAD}$ ($R = 0.81$). Both correlations are statistically significant (p-value < 0.05).

3.2.5. Pathway Analysis

For the lung cancer example, we performed the same steps as in Section 3.1.6. Figure 5 reports the networks of not-isolated gene signatures selected by the three screening approaches: (panel A) BMD + DAD, (panel B) BMD, and (panel C) DAD.

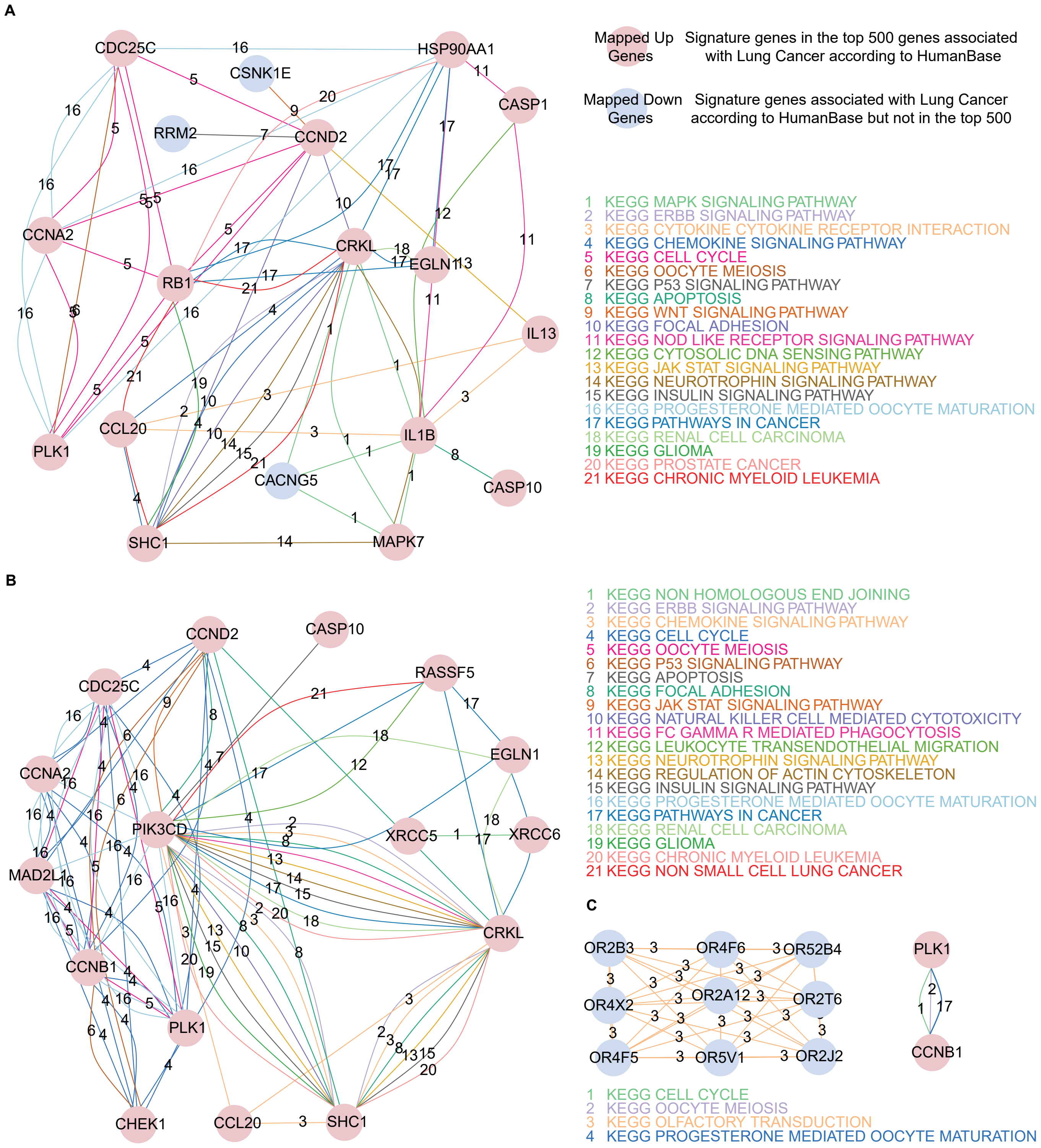
A
Mapped Up Genes
Signature genes in the top 500 genes associated with Lung Cancer according to HumanBase
Mapped Down Genes
Signature genes associated with Lung Cancer according to HumanBase but not in the top 500
CDC25C
HSP90AA1
CSNK1E
CASP1
RRM2
CCND2
CCNA2
CRKL
EGLN1
RB1
IL13
CCL20
IL1B
PLK1
CACNG5
CASP10
SHC1
MAPK7
1 KEGG MAPK SIGNALING PATHWAY
2 KEGG ERBB SIGNALING PATHWAY
3 KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
4 KEGG CHEMOKINE SIGNALING PATHWAY
5 KEGG CELL CYCLE
6 KEGG OOCYTE MEIOSIS
7 KEGG P53 SIGNALING PATHWAY
8 KEGG APOPTOSIS
9 KEGG WNT SIGNALING PATHWAY
10 KEGG FOCAL ADHESION
11 KEGG NOD LIKE RECEPTOR SIGNALING PATHWAY
12 KEGG CYTOSOLIC DNA SENSING PATHWAY
13 KEGG JAK STAT SIGNALING PATHWAY
14 KEGG NEUROTROPHIN SIGNALING PATHWAY
15 KEGG INSULIN SIGNALING PATHWAY
16 KEGG PROGESTERONE MEDIATED OOCYTE MATURATION
17 KEGG PATHWAYS IN CANCER
18 KEGG RENAL CELL CARCINOMA
19 KEGG GLIOMA
20 KEGG PROSTATE CANCER
21 KEGG CHRONIC MYELOID LEUKEMIA
B
CCND2
CASP10
RASSF5
CDC25C
EGLN1
CCNA2
PIK3CD
XRCC5
XRCC6
MAD2L1
CRKL
CCNB1
PLK1
CHEK1
CCL20
SHC1
1 KEGG NON HOMOLOGOUS END JOINING
2 KEGG ERBB SIGNALING PATHWAY
3 KEGG CHEMOKINE SIGNALING PATHWAY
4 KEGG CELL CYCLE
5 KEGG OOCYTE MEIOSIS
6 KEGG P53 SIGNALING PATHWAY
7 KEGG APOPTOSIS
8 KEGG FOCAL ADHESION
9 KEGG JAK STAT SIGNALING PATHWAY
10 KEGG NATURAL KILLER CELL MEDIATED CYTOTOXICITY
11 KEGG FC GAMMA R MEDIATED PHAGOCYTOSIS
12 KEGG LEUKOCYTE TRANSENDOTHELIAL MIGRATION
13 KEGG NEUROTROPHIN SIGNALING PATHWAY
14 KEGG REGULATION OF ACTIN CYTOSKELETON
15 KEGG INSULIN SIGNALING PATHWAY
16 KEGG PROGESTERONE MEDIATED OOCYTE MATURATION
17 KEGG PATHWAYS IN CANCER
18 KEGG RENAL CELL CARCINOMA
19 KEGG GLIOMA
20 KEGG CHRONIC MYELOID LEUKEMIA
21 KEGG NON SMALL CELL LUNG CANCER
C
OR2B3
OR4F6
OR52B4
PLK1
OR4X2
OR2A12
OR2T6
OR4F5
OR5V1
OR2J2
CCNB1
1 KEGG CELL CYCLE
2 KEGG OOCYTE MEIOSIS
3 KEGG OLFACTORY TRANSDUCTION
4 KEGG PROGESTERONE MEDIATED OOCYTE MATURATION

Figure 5. Lung cancer KEGG pathway analysis. KEGG pathway networks of the not-isolated gene signatures selected by BMD + DAD (**A**), BMD (**B**), and DAD (**C**) screening approaches. Red nodes represent the mapped up genes, i.e., signatures that are among the top 500 genes biologically known to be associated with lung cancer according to HumanBase. Blue nodes represent the mapped down genes, i.e., signatures that are biologically known to be associated with lung cancer, but not among the top 500 genes according to HumanBase. Any gene that is not associated with lung cancer will be displayed as a white node (none were selected in the not-isolated gene signatures shown in the figure). `COSMONET` generates the KEGG pathway networks as part of a fully interactive dashboard, which also contains the full networks comprising all the gene signatures including the isolated genes. The full dashboards are available in Appendix A.

The mapped up gene PLK1 was selected by all methods, confirming its key role in lung cancer [53–55]. The BMD + DAD and BMD screening methods (Figure 5 (panel A) and (panel B)) shared seven biologically relevant genes (mapped up genes), including SHC1 and CRKL. The association of these genes with lung cancer is already known [56]. However, by investigating the KEGG pathways shared by both genes (i.e., links between the two nodes), it will be possible to better understand the interaction of the two genes and their role in lung cancer.

By analyzing the links between the mapped up genes and mapped down genes, it is possible to identify new potential biomarkers, as well as the KEGG pathways associated with them. For example, Figure 5 (panel A) shows that the mapped up gene CCND2 is connected to the mapped down gene RRM2 through the KEGG P53 signaling pathway and to the mapped down gene CSNK1E through the KEGG WNT signaling pathways. This information can be useful to understand how the key gene CCND2 interacts with other genes that have not been identified as highly relevant for lung cancer. The interactive networks dashboard is available in Appendix A.2.

3.3. Computational Complexity

To provide an estimate of the runtime of the proposed methodology, we ran the Cosmonet() function using the breast cancer datasets presented above, using two separate matrices for the training and testing phases. The run times were approximately 30 s for the BMD screening, approximately 10 seconds for the DAD screening, and 40 s for the BMD + DAD screening. When running the same procedure using the LUAD dataset (using one single matrix), the run times were approximately 65 s for the BMD screening, approximately 6 s for the DAD screening, and 50 s for the BMD + DAD screening. As expected, the running time was longer when using the LUAD dataset, as this included the train/test splitting phase and a larger number of samples.

It is worth mentioning that the runtime is mostly affected by the size of the input matrices and by the number of genes selected during the screening process. The execution times were measured using the R library tictoc (https://github.com/collectivemedia/tictoc, accessed on 13 November 2021) and the experiments were run on an Intel Core i5 processor with 8 GB of RAM.

4. Discussion and Conclusions

In this work, we presented COSMONET, a novel R package that merges advanced screening techniques with network-based Cox regression for survival prediction. This package implements a complete pipeline built upon an efficient computational algorithm based on two main features: (i) screening techniques to reduce the feature space p to a moderate scale d and (ii) a network-based Cox regression method to select the high-risk cancer genes among the screened covariates for the survival of patients [34,35]. The variable screening aims to find a subset of input and significant variables associated with cancer patient survival. The network-based Cox regression algorithm, based on augmented and penalized minimization-L_0(APM-L_0) [45], aims to select the most functional genes related to cancer. Moreover, the package is completed by several additional functions to determine the optimal cut-off for the PIs, perform data pre-processing, and visualize and compare the results in several forms. We illustrated the capabilities of COSMONET on two case studies with gene expression data (microarrays and RNA-seq) from the literature. The vignette provides a step-by-step guide for beginner users, making COSMONET appealing for use within the biomedical community in order to plan, develop, and implement a personalized care plan.

The current version of COSMONET works with a single type of omic data. We have used gene expression, although it is possible to use other types of omics profiles. Together with several others (i.e., [35,57,58]), we have shown that integrating multiple omics data types can improve the performance of a method. The approach used in [35] can be easily extended to COSMONET. In the aforementioned proposal, we used MANCIE [59] to correct the gene expression data using the copy number aberrations (CNA), which showed better performance. The same approach can be used with COSMONET when two types of omics are available. However, the approach in MANCIE assumes that one of the omics (i.e., the gene expression) acts as the principal matrix and the second one as a support matrix (i.e., the CNA in [35]). Therefore, the approach does not easily extend to multiple omics and does not give equal importance to the experiments. To further improve COSMONET for data integration, we should use the multi-task regression learning approaches in a manner similar to [60,61]. Moreover, we have shown that COSMONET runs fast also on large datasets such as those in our case studies. However, due to the increasing complexity of a multi-omics approach, we expect a longer runtime when integrating multiple datasets or when the size of the problems increases. For this reason, future improvements of the package should integrate the option to run the process in parallel mode. Furthermore, the integration of different databases other than KEGG, such as MetaCyc [62], could also improve the user experience, allowing an analysis of the data from different perspectives. In fact, MetaCys contains significantly more reactions and pathways than KEGG. Hence, we aim to include additional pathway repositories in a future extension of COSMONET.

In conclusion, to the best of our knowledge, COSMONET is the only R package that combines both biologically driven and data-driven screening techniques within a network-penalized Cox regression model. Hence, the proposed package provides the biomedical community with a valuable and straightforward tool for investigating new cancer biomarkers and predicting survival outcomes while using the most recent statistical techniques.

Supplementary Materials: The following supplementary figures are available online at https://www.mdpi.com/article/10.3390/math9243262/s1. Figure S1: Density plots. (A) Breast cancer data; (B) Lung cancer data. Figure S2: Training phase. (A) Kaplan–Meier curves and (B) distribution of the risk score PI^D using the BMD + DAD gene signatures detected in breast cancer. (C) Kaplan–Meier curves and (D) distribution of the risk score PI^D using the BMD + DAD the BMD + DAD gene signatures discovered in lung cancer. For both cases, the p-value is statistically significant (p-value < 0.05). The red line (high-risk, i.e., poor survival) and the blue line (low-risk, i.e., high survival) indicate the two risk conditions associated with cancer survival patients of the testing set D. The risk table shows the number of individuals at risk during the time of follow-up. Figure S3: Heat-maps. (A) BMD + DAD genes signatures in the training set (T) and the testing set (D) selected using breast cancer data. (B) BMD + DAD gene signatures in the training set (T) and the testing set (D) identified using lung cancer data. The genes in the horizontal direction are clustered in the same order for both sets (T and D). Z-scores of the gene expression data are used. Z-scores larger than 3.5 were set to 3.5 and Z-scores smaller than -3.5 were set to -3.5. Red color indicates a high level of expression in breast cancer and green color indicates a low level of expression. The patients are divided into high-risk and low-risk groups based on the optimal cut-off $PI^{T,*}$.

Author Contributions: Conceptualization, A.I. and A.O.; methodology, A.I. and A.O.; software, A.I. and A.O.; validation, A.I. and A.O.; formal analysis, A.I. and A.O.; investigation, A.I. and A.O.; resources, A.I. and A.O.; data curation, A.I. and A.O.; writing—original draft preparation, A.I., A.O., C.A., I.D.F. and P.L.; writing—review and editing, A.I., A.O., C.A., I.D.F. and P.L.; visualization, A.I., A.O., C.A., I.D.F. and P.L.; supervision, C.A., I.D.F. and P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the project "Antitumor Drugs and Vaccines from the Sea (ADViSE)" project (CUP B43D18000240007–SURF 17061BP000000011) funded by POR Campania FESR 2014-2020 "Technology Platform for Therapeutic Strategies against Cancer"—Action 1.2.1 and 1.2.2. P.L. is funded by the EU Horizon 2020 GO-DS21 (grant agreement No 848077). A.O. would like to thank the support from Earlier.org (grant agreement No 11215056).

Institutional Review Board Statement: Not applicable since publicly available data were used.

Informed Consent Statement: Not applicable.

Data Availability Statement: The GEO datasets analyzed during the current study are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2990 (accessed on 5 June 2021). The TCGA-LUAD data analyzed during the current study are available at http://linkedomics.org/data_download/TCGA-LUAD/ (accessed on 5 June 2021). The Disease Ontology (DO) can be linked via the Disease Ontology ID at https://www.disease-ontology.org/?id=DOID:1612 for breast cancer (accessed on 5 June 2021) and https://www.disease-ontology.org/?id=DOID:1324 for lung cancer (accessed on 5 June 2021). `COSMONET` R package is available at http://bioinfo.na.iac.cnr.it/cosmonet/ (accessed on 5 June 2021). The step-by-step R vignette can be accessed at http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadVignette/vignette/vignettes.html on 13 November 2021.

Acknowledgments: This study made use of data collected in the GEO database and GDC data portal.

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

Abbreviations

The following abbreviations are used in this manuscript:

n	sample size
p	future space
T	training set
n_T	training sample size
D	testing set
n_D	testing sample size
$\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip})^T$	covariates of the ith sample over p
$Y_i = min(t, c_i)$	response variable (t_i is the survival time and c_i the censoring time)
$\delta_i = I(t_i < c_i)$	censoring indicator ($I(\cdot)$ indicator function)
BMD screening	Biomedical screening
DAD screening	Data-driven screening
BMD + DAD screening	union of BMD and DAD screening
$d < p$	screening future space
$\mathcal{I}_{BMD}$	screening set variables for BMD screening
$\mathcal{I}_{DAD}$	screening set variables for DAD screening
$\mathcal{I}_{BMD+DAD}$	screening set variables for BMD + DAD screening
$\boldsymbol{\beta}_{\mathcal{I}} = (\beta_1, \ldots, \beta_d)^T$	regression coefficients
α, λ and ρ	regularization parameters
PI^T	prognostic index on T for each screening (PI^T_{BMD}, PI^T_{DAD}, $PI^T_{BMD+DAD}$)
$q_\gamma = (q_{\gamma_1}, \ldots, q_{\gamma_j})$	sample quantiles based on:
	adaptive-based approach: $(\gamma_1, \ldots, \gamma_j) = (0.25, \ldots, 0.80)$
	median-based approach: $\gamma = 0.50$
	survival-based approach: $\gamma = \xi$ where ξ is computed using the function `surv_cutpoint` from `survminer` package
$PI^{T,*}$	optimal cut-off on T ($PI^{T,*}_{BMD}$, $PI^{T,*}_{DAD}$, $PI^{T,*}_{BMD+DAD}$)
PI^D	prognostic index on D (PI^D_{BMD}, PI^D_{DAD}, $PI^D_{BMD+DAD}$)
$n_{T_{HR}}$ and $n_{D_{HR}}$	number of patients at high-risk in T and D, respectively
$n_{T_{LR}}$ and $n_{D_{LR}}$	number of patients at low-risk in T and D, respectively
r	correlation coefficient between prognostic indices PIs

Appendix A. Pathway Analysis Dashboard

Interactive dashboards can be generated by using `COSMONET`. Some examples are available at the following links. The dashboard contains two tabs, one for the not-isolated genes and one for the full network. By hovering over the networks' edges, the name of the corresponding pathway is shown.

Appendix A.1. Breast Cancer Data Dashboards

- http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadDashboard/dashboard/RUNCosmonetDashboardBreast_BMD.html (accessed on 13 November 2021)

- http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadDashboard/dashboard/RUNCosmonetDashboardBreast_DAD.html (accessed on 13 November 2021)
- http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadDashboard/dashboard/RUNCosmonetDashboardBreast_BMD_DAD.html (accessed on 13 November 2021)

Appendix A.2. Lung Cancer Data Dashboards

- http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadDashboard/dashboard/RUNCosmonetDashboardLung_BMD.html (accessed on 13 November 2021)
- http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadDashboard/dashboard/RUNCosmonetDashboardLung_DAD.html (accessed on 13 November 2021)
- http://bioinfo.na.iac.cnr.it/cosmonet/cosmonet/downloadDashboard/dashboard/RUNCosmonetDashboardLung_BMD_DAD.html (accessed on 13 November 2021)

References

1. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a Shared Vision for Cancer Genomic Data. 2016. Available online: https://portal.gdc.cancer.gov (accessed on 5 June 2021).
2. Gao, G.F.; Parker, J.S.; Reynolds, S.M.; Silva, T.C.; Wang, L.B.; Zhou, W.; Akbani, R.; Bailey, M.; Balu, S.; Berman, B.P.; et al. Before and after: Comparison of legacy and harmonized TCGA genomic data commons' data. *Cell Syst.* **2019**, *9*, 24–34. [CrossRef] [PubMed]
3. Lappalainen, I.; Almeida-King, J.; Kumanduri, V.; Senf, A.; Spalding, J.D.; Saunders, G.; Kandasamy, J.; Caccamo, M.; Leinonen, R.; Vaughan, B.; et al. The European Genome-Phenome Archive of Human Data Consented for Biomedical Research. 2015. Available online: http://www.ebi.ac.uk/ega/ (accessed on 5 June 2021).
4. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for Functional Genomics Data Sets—Update. 2012. Available online: https://www.ncbi.nlm.nih.gov/geo/ (accessed on 5 June 2021).
5. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef]
6. Herrmann, M.; Probst, P.; Hornung, R.; Jurinovic, V.; Boulesteix, A.L. Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief. Bioinform.* **2021**, *22*, bbaa167. [CrossRef] [PubMed]
7. Thapa, S.; Lomholt, M.A.; Krog, J.; Cherstvy, A.G.; Metzler, R. Bayesian analysis of single-particle tracking data using the nested-sampling algorithm: Maximum-likelihood model selection applied to stochastic-diffusivity data. *Phys. Chem. Chem. Phys.* **2018**, *20*, 29018–29037. [CrossRef]
8. Muñoz-Gil, G.; Garcia-March, M.A.; Manzo, C.; Martín-Guerrero, J.D.; Lewenstein, M. Single trajectory characterization via machine learning. *New J. Phys.* **2020**, *22*, 013010. [CrossRef]
9. Malik, M.A.; Malik, S.A.; Haq, M.G.; Bangri, S.A.; Ahmad, S.Z.; Shah, O.J.; Shah, Z.A. Genetic Susceptibility of DCC Gene in Gallbladder Cancer in Kashmir and Meta-Analysis. *Nutr. Cancer* **2021**, 1–9. [CrossRef] [PubMed]
10. Vinga, S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief. Bioinform.* **2020**, *22*, 77–87. [CrossRef]
11. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **1997**, *16*, 385–395. [CrossRef]
12. Gui, J.; Li, H. Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics* **2005**, *21*, 3001–3008. [CrossRef]
13. Goeman, J.J. L1 penalized estimation in the Cox proportional hazards model. *Biom. J.* **2010**, *52*, 70–84.
14. Engler, D.; Li, Y. Survival analysis with high-dimensional covariates: An application in microarray studies. *Stat. Appl. Genet. Mol. Biol.* **2009**, *8*, 14. [CrossRef]
15. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **2011**, *39*, 1–13. [CrossRef] [PubMed]
16. Fan, J.; Li, R. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.* **2002**, *30*, 74–99. [CrossRef]
17. Zhang, H.H.; Lu, W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **2007**, *94*, 691–703. [CrossRef]
18. Antoniadis, A.; Fryzlewicz, P.; Frederique, L. The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **2011**, *37*, 531–552. [CrossRef]
19. Li, C.; Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **2008**, *24*, 1175–1182. [CrossRef] [PubMed]
20. Li, C.; Li, H. Variable selection and regression analysis for graph structured covariates with an application to genomics. *Ann. Appl. Stat.* **2010**, *4*, 1498–1516. [CrossRef]
21. Wu, T.T.; Wang, S. Doubly Regularized Cox Regression for High-dimensional Survival Data with Group Structures. *Stat. Its Interface* **2013**, *6*, 175–186.

22. Gong, H.; Wu, T.T.; Clarke, E.M. Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression. *BMC Syst. Biol.* **2014**, *8*, 1–9. [CrossRef]
23. Sun, H.; Lin, W.; Feng, R.; Li, H. Network-regularized high-dimensional Cox regression for analysis of genomic data. *Stat. Sin.* **2014**, *24*, 1433. [CrossRef]
24. Zhang, W.; Ota, T.; Shridhar, V.; Chien, J.; Wu, B.; Kuang, R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol.* **2013**, *9*, e1002975. [CrossRef] [PubMed]
25. Jiang, H.K.; Liang, Y. The L1/2 regularization network Cox model for analysis of genomic data. *Comput. Biol. Med.* **2018**, *100*, 203–208. [CrossRef]
26. Veríssimo, A.; Oliveira, A.L.; Sagot, M.F.; Vinga, S. DegreeCox—A network-based regularization method for survival analysis. *BMC Bioinform.* **2016**, *17*, 449. [CrossRef] [PubMed]
27. Demchak, B.; Kreisberg, J.F.; Bass, J.I.F. Theory and Application of Network Biology Toward Precision Medicine. *J. Mol. Biol.* **2018**, *430*, 18 Pt A, 2873–2874. [CrossRef]
28. Zhang, W.; Chien, J.; Yong, J.; Kuang, R. Network-based machine learning and graph theory algorithms for precision oncology. *NPJ Precis. Oncol.* **2017**, *1*, 25. [CrossRef]
29. Zhao, Y.; Chang, C.; Long, Q. Knowledge-Guided Statistical Learning Methods for Analysis of High-Dimensional -Omics Data in Precision Oncology. *JCP Precis. Oncol.* **2019**, *3*, 1–9. [CrossRef]
30. Ozturk, K.; Dow, M.; Carlin, D.E.; Bejar, R.; Carter, H. The emerging potential for network analysis to inform precision cancer medicine. *J. Mol. Biol.* **2018**, *430*, 2875–2899. [CrossRef]
31. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
32. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
33. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef] [PubMed]
34. Iuliano, A.; Occhipinti, A.; Angelini, C.; De Feis, I.; Lió, P. Cancer Markers Selection Using Network-Based Cox Regression: A Methodological and Computational Practice. *Front. Physiol.* **2016**, *7*, 208. [CrossRef] [PubMed]
35. Iuliano, A.; Occhipinti, A.; Angelini, C.; De Feis, I.; Liò, P. Combining pathway identification and breast cancer survival prediction via screening-network methods. *Front. Genet.* **2018**, *9*, 206. [CrossRef] [PubMed]
36. Veríssimo, A.; Carrasquinha, E.; Lopes, M.B.; Oliveira, A.L.; Sagot, M.F.; Vinga, S. Sparse Network-Based Regularization for the Analysis of Patientomics High-Dimensional Survival Data. 2018. Available online: https://www.bioconductor.org/packages/release/bioc/html/glmSparseNet.html (accessed on 5 June 2021).
37. Tissier, R. GitHub Repository. Available online: https://github.com/RenTissier/NetPred.git (accessed on 5 June 2018).
38. Tissier, R.; Houwing-Duistermaat, J.; Rodríguez-Girondo, M. Improving stability of prediction models based on correlated omics data by using network approaches. *PLoS ONE* **2018**, *13*, e0192853. [CrossRef] [PubMed]
39. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 849–911. [CrossRef] [PubMed]
40. Fan, J.; Song, R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Stat.* **2010**, *38*, 3567–3604. [CrossRef]
41. Huttenhower, C.; Haley, E.M.; Hibbs, M.A.; Dumeaux, V.; Barrett, D.R.; Coller, H.A.; Troyanskaya, O.G. Exploring the human genome with functional maps. *Genome Res.* **2009**, *19*, 1093–1106. [CrossRef] [PubMed]
42. HumanBase: Data-Driven Predictions of Gene Expression, Function, Regulation, and Interactions in Human. Available online: https://hb.flatironinstitute.org (accessed on 5 June 2021).
43. Fan, J.; Feng, Y.; Wu, Y. High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*; Institute of Mathematical Statistics: Ann Arbor, MI, USA, 2010; pp. 70–86.
44. Greene, C.; Krishnan, A.; Wong, A.; Ricciotti, E.; Zelaya, R.; Himmelstein, D.; Zhang, R.; Hartmann, B.; Zaslavsky, E.; Sealfon, S.; et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **2015**, *47*, 569–576. [CrossRef] [PubMed]
45. Li, X.; Xie, S.; Zeng, D.; Wang, Y. Efficient L 0-norm feature selection based on augmented and penalized minimization. *Stat. Med.* **2018**, *37*, 473–486. [CrossRef]
46. Hastie, T.; Qian, J. Glmnet vignette. *Retrieved June* **2014**, *9*, 1–30.
47. Schriml, L.M.; Mitraka, E.; Munro, J.; Tauber, B.; Schor, M.; Nickle, L.; Felix, V.; Jeng, L.; Bearer, C.; Lichenstein, R.; et al. Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Res.* **2019**, *47*, D955–D962. [CrossRef]
48. Wang, Y.; Klijn, J.G.; Zhang, Y.; Sieuwerts, A.M.; Look, M.P.; Yang, F.; Talantov, D.; Timmermans, M.; Meijer-van Gelder, M.E.; Yu, J.; et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **2005**, *365*, 671–679. [CrossRef]

49. Sotiriou, C.; Wirapati, P.; Loi, S.; Harris, A.; Fox, S.; Smeds, J.; Nordgren, H.; Farmer, P.; Praz, V.; Haibe-Kains, B.; et al. Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **2006**, *98*, 262–272. [CrossRef]
50. Vasaikar, S.V.; Straub, P.; Wang, J.; Zhang, B. LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **2018**, *46*, D956–D963. [CrossRef]
51. Korkaya, H.; Liu, S.; Wicha, M.S. Breast cancer stem cells, cytokine networks, and the tumor microenvironment. *J. Clin. Investig.* **2011**, *121*, 3804–3809. [CrossRef]
52. Yan, L.; Anderson, G.M.; DeWitte, M.; Nakada, M.T. Therapeutic potential of cytokine and chemokine antagonists in cancer therapy. *Eur. J. Cancer* **2006**, *42*, 793–802. [CrossRef] [PubMed]
53. Reda, M.; Ngamcherdtrakul, W.; Gu, S.; Bejan, D.S.; Siriwon, N.; Gray, J.W.; Yantasee, W. PLK1 and EGFR targeted nanoparticle as a radiation sensitizer for non-small cell lung cancer. *Cancer Lett.* **2019**, *467*, 9–18. [CrossRef]
54. Shin, S.B.; Jang, H.R.; Xu, R.; Won, J.Y.; Yim, H. Active PLK1-driven metastasis is amplified by TGF-β signaling that forms a positive feedback loop in non-small cell lung cancer. *Oncogene* **2020**, *39*, 767–785. [CrossRef] [PubMed]
55. Allera-Moreau, C.; Rouquette, I.; Lepage, B.A.; Oumouhou, N.; Walschaerts, M.; Leconte, E.; Schilling, V.; Gordien, K.; Brouchet, L.; Delisle, M.; et al. DNA replication stress response involving PLK1, CDC6, POLQ, RAD51 and CLASPIN upregulation prognoses the outcome of early/mid-stage non-small cell lung cancer patients. *Oncogenesis* **2012**, *1*, e30. [CrossRef] [PubMed]
56. Liang, Y.; Lei, Y.; Du, M.; Liang, M.; Liu, Z.; Li, X.; Gao, Y. The increased expression and aberrant methylation of SHC1 in non–small cell lung cancer: Integrative analysis of clinical and bioinformatics databases. *J. Cell. Mol. Med.* **2021**, *25*, 7039–7051. [CrossRef] [PubMed]
57. Zhu, R.; Zhao, Q.; Zhao, H.; Ma, S. Integrating multidimensional omics data for cancer outcome. *Biostatistics* **2016**, *17*, 605–618. [CrossRef] [PubMed]
58. Pineda, S.; Real, F.X.; Kogevinas, M.; Carrato, A.; Chanock, S.J.; Malats, N.; Van Steen, K. Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. *PLoS Genet.* **2015**, *11*, e1005689. [CrossRef] [PubMed]
59. Zang, C.; Wang, T.; Deng, K. High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* **2016**, *7*, 1–8. [CrossRef] [PubMed]
60. Lin, D.; Zhang, J.; Li, J.; He, H.; Deng, H.W.; Wang, Y.P. Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. *Front. Cell Dev. Biol.* **2014**, *2*, 62. [CrossRef] [PubMed]
61. Cao, H.; Zhou, J.; Schwarz, E. RMTL: An R library for multi-task learning. *Bioinformatics* **2019**, *35*, 1797–1798. [CrossRef] [PubMed]
62. Caspi, R.; Altman, T.; Dale, J.M.; Dreher, K.; Fulcher, C.A.; Gilham, F.; Kaipa, P.; Karthikeyan, A.S.; Kothari, A.; Krummenacker, M.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2010**, *38*, D473–D479. [CrossRef] [PubMed]