Cancer Markers Selection Using Network-Based Cox Regression: A Methodological and Computational Practice

Antonella Iuliano¹*†, Annalisa Occhipinti²*†, Claudia Angelini¹, Italia De Feis¹ and Pietro Lió²

¹Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Naples, Italy; ²Computer Laboratory, University of Cambridge, Cambridge, UK

International initiatives such as the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are collecting multiple datasets at different genome-scales with the aim of identifying novel cancer biomarkers and predicting survival of patients. To analyze such data, several statistical methods have been applied, among them Cox regression models. Although these models provide a good statistical framework to analyze omic data, there is still a lack of studies that illustrate advantages and drawbacks in integrating biological information and selecting groups of biomarkers. In fact, classical Cox regression algorithms focus on the selection of a single biomarker, without taking into account the strong correlation between genes. Even though network-based Cox regression algorithms overcome such drawbacks, such network-based approaches are less widely used within the life science community. In this article, we aim to provide a clear methodological framework on the use of such approaches in order to turn cancer research results into clinical applications. Therefore, we first discuss the rationale and the practical usage of three recently proposed network-based Cox regression algorithms (i.e., Net-Cox, AdaLnet, and fastcox). Then, we show how to combine existing biological knowledge and available data with such algorithms to identify networks of cancer biomarkers and to estimate survival of patients. Finally, we describe in detail a new permutation-based approach to better validate the significance of the selection in terms of cancer gene signatures and pathway/networks identification. We illustrate the proposed methodology by means of both simulations and real case studies. Overall, the aim of our work is two-fold. Firstly, to show how network-based Cox regression models can be used to integrate biological knowledge (e.g., multi-omics data) for the analysis of survival data. Secondly, to provide a clear methodological and computational approach for investigating cancers regulatory networks.

Keywords: cancer, Cox model, high-dimensionality, gene expression, network, regularization, survival
INTRODUCTION

Recent developments in high-throughput technology have produced a huge amount of multiple and diverse genome-scale data to deal with biological and clinical questions in cancer. For example, genomics, transcriptomics, and epigenomics information is nowadays publicly available for tens of different cancer cell lines from thousands of patients in The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). Mutations data over one million tumor samples are also reported in Cosmic (http://cancer.sanger.ac.uk/cosmic), the world’s largest and most comprehensive resource for exploring the impact of somatic mutations. Other valuable databases include The Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/gds) among others. Such amount of data is likely to revolutionize genetics and biomedical cancer research, but a thorough integration of all these different types of information is necessary. Indeed, cancer is a “multi-factorial” disease caused by a combination of genetic, environmental, and lifestyle factors. Such factors play an important role in discovering prognostic and diagnostic cancer gene signatures opening a new way toward the so called “personalized medicine.” The term refers to a new type of therapy that is essentially based on the features of each patient. For instance, the anticancer drug Cetuximab (Karapetis et al., 2008) inhibits cells proliferation by binding to the EGF receptor and, consequently, preventing activation of the downstream signaling pathway. However, it has been found that Cetuximab can work only if the K-RAS gene is not mutated. Another example is the anti-cancer drug Trastuzumab (Hudis, 2007), which is effective only in patients that highly express the human epidermal growth factor (HER2) at the cell surface, to which the antibody binds. These examples highlight the need of identifying stable and interpretable biomarkers able to predict patient survival and characterize a patient-personalized therapy. In addition, the knowledge of complex cancer processes and networks is important to optimize the use of technology within health care (Raghupathi and Raghupathi, 2014). By discovering associations within the data, big data analytics has the potential to improve care, save lives, and lower costs.

As a consequence, in the last years, there has been a growing interest in developing methods that integrate different genome-scale data into regression models for survival data to create a comprehensive view of human biology and disease (Wang et al., 2014). A popular used approach for the integration of genomic and clinical information is the Cox proportional hazard model (Cox, 1972). The main goal of such method is investigating the connection between gene expression data and survival information to predict cancer survival, assess cancer outcomes, and identify new gene markers. However, since gene expression data are usually characterized by a number of covariates $p$ much larger than the sample size $n$, the traditional Cox model cannot be applied. Hence, several penalized Cox regression methods have been developed to identify core pathways and biomarkers involved in cancer progression, e.g., the Cox model based on Lasso penalty (Tibshirani, 1996, 1997; Gui and Li, 2005). Alternative penalized Cox regression models based on variable selection include the SCAD (Fan and Li, 2001), the adaptive Lasso (Zou, 2006), the elastic net model (Zou and Hastie, 2005; Simon et al., 2011a; Wu, 2012), and the Danszig selector (Candes and Tao, 2007) among others. These methods are able to cope with the high-dimensionality of gene expression data, thus solving the $p \gg n$ issue (Engler and Li, 2009). All these penalized models are statistically efficient in high-dimensional regression, but they perform poorly on data with high collinearity. Moreover, no biological knowledge is taken into account. Indeed, they are simply based on statistical frameworks completely ignoring biological regulatory network, protein–protein interaction (PPI), signaling pathways, and well-known relationships among genes. In such models, the lack of biological information produces instability in predictors reducing the predictive ability of the models. Hence, in order to provide more reliable and biologically meaningful results, the inclusion of a-priori biological knowledge into the models is mandatory. To address this issue, new penalized Cox methods based on the integration of genomic information have been recently proposed (Zhang et al., 2013; Gong et al., 2014; Sun et al., 2014). In such models, the genomic information is encoded by a network whose graph structure identifies a given relation (edges) between genes (nodes). The resulting Laplacian matrix is then integrated as penalty in the Cox regression models. In particular, the network can represent the correlation between genes (Zhang et al., 2013), KEGG pathways identification (Sun et al., 2014), functional interaction network (Huttenhower et al., 2009), or PPI. These Cox models based on a-priori biological network are called “network-based Cox regression.”

The network-based Cox regression methods provide an efficient tool to perform Cox regression on high-dimensional data incorporating genes network information. In literature, there are some recent approaches that analyze different Cox methods. For instance, an accurate review of eight different methods that integrate network information into multi-variable Cox models is presented to study the risk prediction in breast cancer and the integrated Brier score is used as a performance measure (Fröhlich, 2014). However, the study performed enrichment analysis on the signatures genes selected by the compared models without showing any survival prediction analysis in terms of Kaplan–Meier curves. A network-based Cox regression model that explores gene-to-gene connections in multiple cancer datasets is also performed for maximizing the overall association of the sub-network with clinical outcomes (Martinez-Ledesma et al., 2015). A potential limitation of these conventional networks is that the edges only reflect the information of within-features or within-relations, and do not consider the association between features and outcomes, which may be useful in improving the predictive power. Therefore, an alternative network construction method for the outcome-guided gene-interaction network has to be introduced in order to improve the performance of survival analysis in network-based Cox regression (Jeong et al., 2015).

In this work, we present a methodological framework for the analysis of molecular and survival data through a cross-validated approach of network-based Cox regression algorithms (Net-Cox, Adalnet, and fastcox, see Section Methods). The method starts from the analysis of raw data and, through a
cross-validated penalty approach, it guides the reader to the interpretation of the final results. As shown in Figure 1, the general steps of our approach are the following: (i) defining the biological question and the experimental design using microarray data, then integrating a-priori biological information using functional map of the human genome such as HEPalmp (Huttenhower et al., 2009) and KEGG; (ii) performing biological screening of the data for selecting relevant features through cross-validated penalization (Simon et al., 2011b); (iii) implementing network-based Cox regression models for the analysis of cancer-related genes; (iv) evaluating survival models to predict cancer patient prognosis and exploring cancer associated pathways. The presented approach provides a new methodological framework for the study and the interpretation of regression methods through gene-network and pathways analyses and it can be easily adapted to incorporate other network-based Cox regression algorithms.

A preliminary study for the comparison of penalized Cox models was presented in Iuliano et al. (2014), where the analysis was limited to cancer survival prediction using top ranked genes. No simulation studies, extensive pathways analysis or validation of the data were performed in that study. On the contrary, this article presents a more accurate and complete analysis based on a cross-validated approach (Simon et al., 2011b), the overall workflow (see Figure 2) that includes both simulation studies and novel real cancer datasets (see Section Data Analysis). Simulated data have been used to perform a statistical comparison of the methods in terms of sensitivity, specificity, number of selected genes, false positive rates, and Matthews correlation coefficient in two simulation settings with different genetic effects. On the other hand, real datasets analysis was performed to assess the relevance of the selected genes in the training dataset and to test the survival prediction accuracy of each model. Cross-validated Kaplan–Meier curves for survival analysis and pathway analysis were also computed (see Section Results). The novelty of the current study consists in the integration of a cross-validated approach (Simon et al., 2011b) to obtain an accurate survival prediction even when the number of cases is relatively small for an effective sample splitting (see Figure 2). Cross-validation methods have been largely applied in Cox regression models to estimate prediction errors and for model parameters tuning (Vasselli et al., 2003; Molinaro et al., 2005; Simon et al., 2011b). Some of the most relevant cross-validation approaches include leave-one-out cross-validation (LOOCV; Kearns and Ron, 1999), k-fold (Refaeilzadeh et al., 2009), and bootstrap algorithms (Kohavi, 1995). However, all these methods do not provide a good estimation if the data available are limited for an effective division in training and test sets. On the contrary, the cross-validation method used in our analysis (Simon et al., 2011b) is based on a re-sampling algorithm that allows an accurate prediction of the survival risk model regardless the data size. Therefore, in this work, we first present a novel statistical approach to infer pathway interaction networks from gene expression data that relies on a new mathematical
FIGURE 2 | Workflow of prognostic model building by using gene expression profile in cancer. The method starts from the analysis of raw data and, through a cross-validated penalty approach, it leads to the interpretation of the final results. Step (1) includes the input data for the survival analysis: gene expression data.
concept (based on the biological screening and network-based Cox regression methods) for understanding pathways’ activity and relationships. Second, we provide a methodological strategy to researchers for the use of network-based Cox regression models in order to turn cancer research results into clinical applications.

**METHODS**

**Network-Regularized Cox Regression Models**

The Cox Proportional hazards model (Cox, 1972) is the most widely used model to describe the relationship between survival times and predictor covariates.

Given a sample of $n$ subjects, let $T_i$ and $C_i$ be the survival time and the censoring time, respectively, for subject $i = 1, \ldots, n$. Let $t_i = \min\{T_i, C_i\}$ be the observed survival time and $\delta_i = I(T_i \leq C_i)$ the censoring indicator, where $I(\cdot)$ is the indicator function (i.e., $\delta_i = 1$ if the survival time is observed and $\delta_i = 0$ if the survival time is censored). We denote by $X_i = (X_{i1}, \ldots, X_{ip})^\prime$ the regression vector of $p$-variables for the $i$th subject (i.e., the gene expression profile of the $i$th patient over $p$ genes). The survival time $T_i$ and the censoring time $C_i$ are assumed to be conditionally independent given $X_i$. Furthermore, the censoring mechanism is assumed to be non-informative. The observed data can be represented by the triplets $\{(t_i, \delta_i, X_i) : i = 1, \ldots, n\}$. The Cox regression method assumes that the hazard function $h(t|X_i)$, which is the risk of death at time $t$ for the $i$th patient with gene expression profile $X_i$, can be written as

$$h(t|X_i) = h_0(t) \exp (X_i^\prime \beta)$$

where $h_0(t)$ is the baseline hazard and $\beta = (\beta_1, \ldots, \beta_p)^\prime$ is the column vector of the regression parameters.

In the classical setting, the regression coefficients are estimated by maximizing the Cox’s log-partial likelihood

$$p_l(\beta) = \sum_{i=1}^{n} \delta_i \left\{X_i^\prime \beta - \log \left[ \sum_{j \in R(t_i)} \exp (X_j^\prime \beta) \right]\right\},$$

where $t_i$ is the survival time (observed or censored) for the $i$th patient, $R(t_i)$ is the risk set at time $t_i$ (i.e., the set of all patients who still survived prior to time $t_i$).

However, in the analysis of gene expression data, the number of genes $p$ is usually larger than the sample size $n$ and the standard Cox-model cannot be directly applied. To cope with the curse of dimensionality ($p \gg n$), a variety of penalization approaches have been proposed for achieving good prediction performance and easy interpretation of the data. Although these regularization methods induce sparsity into the solution by shrinking some estimates to zero, the biological relationship of gene expression profiles is not taken into account. Hence, in order to integrate information from molecular interactions between genes, network-based constrained methods for high-dimensional Cox regression have been introduced.

In this context, the regression coefficients are estimated by maximizing the penalized log-partial likelihood function

$$p_l^{pen}(\beta) = \sum_{i=1}^{n} \delta_i \left\{X_i^\prime \beta - \log \left[ \sum_{j \in R(t_i)} \exp (X_j^\prime \beta) \right]\right\} - P(\beta),$$

where $P(\beta)$ is a network-constrained penalty function on the coefficients $\beta$.

Such penalty function describes the existing relationships among the covariates (genes) specified by a network $G = (V, E, W)$ (weighted and undirected graph), where $V = \{1, \ldots, p\}$ is the set of vertices (genes/covariates), an element $(i, j)$ in the edge set $E \subseteq V \times V$ indicates a link between vertices $i$ and $j$ and $W = (w_{ij}), (i, j) \in E$ is the set of weights associated with these edges. These weights are usually used to represent the relations between genes in terms of gene–gene interaction, KEGG pathway analysis or PPI. Hence, the network structure plays an important role since it incorporates prior gene regulatory information often ignored.

The three regularized network-based Cox regression models used in our study are presented below and differ in the form of the penalty function $P(\beta)$.

**Net-Cox method**

Net-Cox regression (Zhang et al., 2013) is an extension of the $L_2$-Cox model and uses the following penalty function

$$P_{\lambda, \alpha}(\beta) = \lambda \left[ \alpha \|\beta\|_2^2 + (1 - \alpha) \Phi(\beta) \right],$$

where $\lambda > 0$ and $\alpha \in (0, 1]$ are two regularization parameters in the network constraint and

$$\Phi(\beta) = \sum_{(i,j) \in E} w_{ij}(\beta_i - \beta_j)^2.$$
can be also written as $\Phi(\beta) = \beta'(I - W)\beta = \beta'\tilde{L}\beta$ where $\tilde{L}$ is a positive semi-definite matrix derived from network information (weight matrix $W$) and $I$ is an identity matrix. Hence, the objective function will result in a significant cost in the network if any pair of genes is connected by a high weight edge and the difference between their coefficients is large.

Note that to identify the signature genes classified by Net-Cox, which is a ridge regression based method, we create a consensus ranking of the relevant cancer genes.

**AdaLnet Method**

Adaptive Laplacian net (Sun et al., 2014) is a modified version of a network-constrained regularization procedure for fitting linear models and for variable selection (Li and Li, 2008, 2010) where the predictors are genomic data with graphical structures. AdaLnet is based on prior gene regulatory network information, represented by an undirected graph for the analysis of gene expression data and survival outcomes.

Denoting with $d_i = \sum_{(i,j) \in E} w_{ij}$ the degree of vertex $i$, AdaLnet defines the normalized Laplacian matrix $L = (l_{ij})$ of the graph $G$ by

$$l_{i,j} = \begin{cases} 1, & \text{if } i = j \text{ and } d_i \neq 0, \\ -w_{ij}/\sqrt{d_i d_j}, & \text{if } (i,j) \in E, \\ 0, & \text{otherwise}. \end{cases}$$

(5)

Note that $L$ is positive semi definite. The network-constrained penalty in Equation (2) is given by

$$P_{\lambda,\alpha}(\beta) = \lambda \left[ \alpha \|\beta\|_1 + (1 - \alpha) \Psi(\beta) \right],$$

(6)

with

$$\Psi(\beta) = \sum_{(i,j) \in E} w_{ij} \left( \text{sign}(\tilde{\beta}_i)\tilde{\beta}_i/\sqrt{d_i} - \text{sign}(\tilde{\beta}_j)\tilde{\beta}_j/\sqrt{d_j} \right)^2.$$ (7)

Equation (6) is composed by two penalty terms. The first one is an $L_1$-penalty that induces a sparse solution, the second one is a quadratic Laplacian penalty $\Psi(\beta) = \beta'\tilde{L}\beta$ that imposes smoothness of the parameters $\beta$ between neighboring vertices in the network. Note that $\tilde{L} = S'LS$ with $S = \text{diag}(\text{sign}(\tilde{\beta}_1), \ldots, \text{sign}(\tilde{\beta}_p))$ and $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ is obtained from a preliminary regression analysis. The scaling of the coefficients $\tilde{\beta}$ respect to the degree allows the genes with more connections (i.e., the hub genes) to have larger coefficients. Hence, small changes of expression levels of these genes can lead to large changes in the response.

An advantage of using penalty (6) consists in representing the case when two neighboring variables have opposite regression coefficient signs, which is reasonable in network-based analysis of gene expression data. Indeed, when a transcription factor (TF) positively regulate gene $i$ and negatively regulate gene $j$ in a certain pathway, the corresponding coefficients will result with opposite sign.

Note that in Net-Cox and AdaLnet, $\lambda$ is the parameter controlling the weight between the likelihood and the network constraint and $\alpha \in (0, 1]$ is the parameter weighting the network constraint.

**Fastcox Method**

The penalty function of fastcox (Yang and Zou, 2012) computes the solution paths of the elastic net penalized Cox’s proportional hazards model (Wu, 2012). In this method the penalty function in Equation (2) is given by

$$P_{\lambda,\alpha}(\beta) = \lambda \left[ \alpha w\|\beta\|_1 + \frac{1}{2}(1 - \alpha)\|\beta\|^2_2 \right],$$

where the non-negative weights $w$ allow a more flexible estimation. In particular, setting $w_j = 0$ implies no shrinkage and the variable $j$ will be always included in the final model. Default is $1$ for all variables. $\alpha \in (0, 1]$ is the elastic net trade-off. This regularization technique is a combination of the lasso and ridge penalty that produce a sparse model (given by the $L_1$-penalty) with good prediction accuracy, while encouraging a grouping effect. It is worthy to note that this method does not include any gene network information. It has been used in our study to obtain pathways investigation and survival prediction from a relevant method that is simply based on statistical framework.

**Tuning Parameters by Five-Fold Cross-Validation**

For all the methods, we estimated the regularization parameters using cross-validation. Four-folds of data are used to build a model for validation on the fifth fold, cycling through each of the five-folds in turn. Then, the $(\lambda,\alpha)$ pair that minimizes the cross-validation log-partial likelihood (CVPL) are chosen as the optimal parameters. CVPL is defined as

$$CVPL(\lambda, \alpha) = -\frac{1}{n}\sum_{k=1}^{K} \ell_k(\hat{\beta}^{(-k)}(\lambda, \alpha)) - \ell_k(\tilde{\beta}(\lambda, \alpha)),$$

(8)

where $\tilde{\beta}(\lambda, \alpha)$ is the estimate obtained from excluding the $k$th part of the data with a given pair of $(\lambda, \alpha)$, $\ell(\cdot)$ is the Cox log-partial likelihood on all the sample and $\ell(\cdot)$ is the log-partial likelihood when the $k$th fold is left out (van Houwelingen et al., 2006).

**General Algorithm: A Re-Sampling Method for Survival Prediction**

The prediction capabilities of a given method are usually evaluated using a training set to select the markers and a testing set to measure the goodness of the prediction. In several cases training and test sets are obtained splitting a given dataset in two parts. However, findings could be over optimistic depending on the specific split. To further understand the role of the network information in cross-validation and to overcome the drawbacks of investigating only one split, each network-based model was validated with the re-sampling procedure suggested by Simon et al. (2011b). This method is based on a cross-validated estimate of the survival distribution of the risk groups and provide a more efficient use of data than fixed sample splitting (see Figure 2). The steps of the re-sampling algorithm for survival prediction are presented below.
Procedure 1: k-fold Cross-validated Kaplan–Meier survival method

1. The full dataset D is partitioned into K approximately equal parts D1, . . . , DK.
   For each k = 1, . . . , K
2. Set Tk = D − Dk as the training set and Dk as the testing set.
3. Perform network-based Cox regression on Tk and select high-risk cancer genes Gk. Denote the parameter estimate by βTk.
4. Calculate the prognostic index (PI) for each patient ik in Dk as
   \[ P_{ik}^{Tk} = x_{ik} \hat{\beta}_{Tk} \]
   where \( x_{ik} \) is the vector of gene expression value associated to the \( i_k \)-th patient into the k-fold. Each patient \( i_k \) in Dk is assigned into the high/low-risk group if its prognostic index \( P_{ik}^{Tk} \) is above (or below) a fixed threshold \( P_{I}^{Tk} \) defined adaptively on Tk.
5. All the patients classified as low-and-high risk in any of the folds are grouped together and a single Kaplan–Meier curve is computed as the union of the risk groups defined in each fold. The set of predictive genes is selected as the union of \( G_k \), for \( k = 1, . . . , K \).
6. Compute the log-rank \( \chi^2 \) statistic under the null hypothesis that survival is independent of expression profile.
7. Calculate a permutation p-value as follows:
   \[ \hat{p} = M^{-1} \sum_{i=1}^{M} I(P_m \geq P_0). \]

For our analysis, the estimate \( \hat{\beta}_{Tk} \) in step 4 was computed by using five-fold cross-validation (i.e., \( K = 5 \)) to select the optimal tuning parameter values \( (\hat{\lambda}_{Tk}, \hat{\alpha}_{Tk}) \), that we used to fit the corresponding penalized function \( P_{ik}^{Tk} \) on Tk. In particular, we first set \( \alpha \) to a sufficiently fine grid of values on \([0, 1]\). For each fixed \( \alpha \), \( \lambda \) was chosen from \( \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\} \) for Net-Cox, while it was set \( \lambda \) to a decreasing sequence of values \( \lambda_{\text{max}} \) to \( \lambda_{\text{min}} \) automatically chosen for AdaLnet and fastcox.

In step 5, we selected \( P_{I}^{Tk} \) as the optimal cut-off in terms of \( P_{ik}^{Tk} \). By using the \( P_{ik}^{Tk} \), it was possible to split the patients in two subgroups, i.e., high-risk and low-risk prognosis groups. Thus, the patient \( i_k \) in Tk was assigned to the high-risk (or low-risk) group if his prognostic index \( P_{ik}^{Tk} \) was above (or below) the quantile selected on a grid of given values that spans from 30 to 70%. The cut-off \( P_{I}^{Tk} \) was chosen in correspondence to the lowest \( p \)-value in a log rank test on this grid.

In step 7, we set \( M \) equal to 500.

Survival Analysis

Network-based Cox regression model was used to discover significant variables, i.e., genes, correlated with death risk. Overall survival (OS) curves were estimated using the Cross Kaplan–Meier estimator and compared using the two-sided log-rank test as implemented in the R package survival. The statistical significance of the log-rank statistic related to the cross-validated Kaplan–Meier curves was obtained through a permutation distribution (Simon et al., 2011b) as described in the previous section. Permutation test was used to test the association between high-risk or low-risk groups and \( p < 0.05 \) were considered statistically significant. A simple scheme of the applied procedure for OS estimation is reported in Figure 2.

Furthermore, we also validated the predictive performance of the three methods using independent dataset for training and testing. In this context, we used the largest dataset as training set to identify the gene expression signatures (see Figure 2, step 2). Then, the second independent dataset was considered as test set in order to analyze the survival prediction of the models. We used Kaplan–Meier survival curves and log-rank test to perform the analysis (see Figure 2, step 3).

Pathway Analysis

We performed pathway analysis based on KEGG database and on the Human Experimental/Functional Mapper (Huttenhower et al., 2009). In particular, we focused on a gene–gene interaction analysis developing gene-networks that describe the relations between genes in terms of KEGG pathways. Each node in the network represents a gene and an edge between two nodes means that the two genes belongs to the same pathway. Different colors are used for different pathways. The color of each node indicates how strong is the relationship between the gene and the disease under analysis (ovarian and breast cancer; Huttenhower et al., 2009). The \( p \)-value chosen within the interval \([0, 0.1]\) represents the node color intensity. Red color, that is \( p = 0 \), means that there is a high significant gene-disease relation, while green color, that is \( p = 0.1 \), means that not exist a relevant gene-disease relation.

Gene networks have been computed by considering only the not isolated genes in the intersection between KEGG pathways and the set of genes selected by each method. Given a set of genes \( G \) and the set of all the KEGG pathways \( K \), we defined a gene \( g \) as not isolated if \( G \cap K \supseteq \{g\} \). Namely, \( g \) is not isolated if there is at least another gene \( g' \in G \) belonging to the same pathways of \( g \).

Software

The methodological approach presented in Figure 2 has been implemented as an integrative R script that allows to run the different algorithms under the same R environment. Net-Cox, which is a Matlab toolbox (http://compbio.cs.umn.edu/Net-Cox/), AdaLnet, available as an R code and sent us upon request and fastcox, which is an R package (http://code.google.com/p/fastcox/) were merged together by using R.matlab, https://cran.r-project.org/web/packages/R.matlab/index.html. The script also includes the implementation of the re-sampling permutation approach (Simon et al., 2011b) and the cross-validation method for parameters estimation. Both simulated and real data can
be used to run the script which can be easily adapted for the integration of new Cox models.

For real data analysis, the microarray data were preprocessed using R packages available in Bioconductor. First, we selected from the initial dataset the genes that were more likely to be involved in cancer by using a functional map summarizing the most relevant interactions in the cancer area of interest (Huttenhower et al., 2009). Then, we used HEFalMp tool (Huttenhower et al., 2009) to build the genes network and identify the weight of the edges between the selected genes. Finally, Net-Cox, AdaLnet, and fastcox were implemented integrating a cross-validation method for selecting the optimal tuning parameters $\lambda$ and $\alpha$ and a re-sampling based procedure (Simon et al., 2011b), see Procedure 1.

The scripts are available upon request from the first two authors.

**DATA ANALYSIS**

**Simulation Scheme**

We used the three methods in two different simulation settings (Wu and Wang, 2013; Sun et al., 2014) in order to investigate the performances and the properties of the three models and to facilitate the interpretation of results. We considered two scenarios that are likely to be encountered in genomic studies and we simulated gene expression data as network constrained. Both the two settings consist of 100 regulatory networks. Each regulatory network is composed by one transcription factor (TF) that regulates 10 genes resulting in a total of 1100 genes. Detailed settings are given below.

**Scenario 1: Not-Overlapped Networks**

The first setting simulates a scenario with not-overlapped networks, which means that the 100 regulatory networks are disjoint each other and each gene is linked to only one TF. Under this assumptions, the degree $d_i$ of each TF is 10 and $d_i = 1$ for the regulated genes. The edges’ weight $w_{ij}$ is 1 between the TFs and their regulated genes, $w_{ij} = 0$ otherwise. The expression value of each TF was generated from a normal standard distribution. The expression values of the ten regulated genes were generated from a conditional normal distribution with positive correlation ($\rho = 0.7$) between the expression of five genes and the corresponding TF, and negative correlation ($\rho = -0.7$) for the remaining five genes. This simulates the activation or repression of each gene under the effect of the corresponding TF. The failure times were generated from the Cox model

$$
\lambda(t|X) = \lambda_0(t) \exp\left(\sum_{j=1}^{88} \beta_j X_j\right)
$$

which includes only $s = 88$ relevant genes (i.e., eight regulatory networks). The baseline hazard function $\lambda_0(t)$ was specified by a Weibull distribution with shape parameter 5 and scale parameter 2. Censoring times were generated from $U(2, 15)$ with a censoring rate of about 30%. The sample size was fixed at $n = 200$ and the simulation were replicated 100 times. In this setting of not-overlapped genes, the coefficients $\beta_j, j = 1, \ldots, 44$ were generated from the uniform distribution $U(0.1, 1)$, while $\beta_j, j = 45, \ldots, 88$ were generated from $U(-1.5, -0.1)$.

For each of the settings above, we quantified the noise as the mean between the variance of each transcription factor (TF) and the variance of the 10 corresponding regulated genes.

**Scenario 2: Overlapped Networks**

The second setting simulates a scenario with overlapped networks, where four regulatory networks (i.e., 44 genes) are connected to the other four networks. This mimics the fact that some genes can belong to different pathways regulating different biological processes, as often observed in cancer. For the sake of simplicity, we assume that all the genes (including the TF) in the networks $P_3$, $P_4$, $P_5$, and $P_6$ are connected to the genes in the remaining four network $P_1$, $P_2$, $P_7$, and $P_8$ which are maintained disjointed and independent each other. The expression values of the TFs and the regulated genes were generated from a multivariate normal distribution with $\text{cov}(X_i, X_j) = 0.5|i-j|$. The coefficients $\beta_j, j = 1, \ldots, 22$, corresponding to $P_1$ and $P_2$, were generated from the uniform distribution $U(0.1, 0.5)$, the coefficients corresponding to the 44 common genes $\beta_j, j = 23, \ldots, 66$ were generated from $U(-0.1, 0.1)$ and the coefficients $\beta_j, j = 67, \ldots, 88$, corresponding to $P_7$ and $P_8$, were generated from the uniform distribution $U(-1, -0.5)$. Survival times were generated as reported in the first setting with the same censoring rate.

**Statistical Measures**

The performance of each method is summarized by four measures: sensitivity, specificity, number of genes selected, and the Matthews correlation coefficient (MCC). The sensitivity or true positive rate (TPR) and specificity or true negative rate (TNR) are given by

$$
TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP},
$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. A test with high sensitivity (few false negative) has a low type II error rate, while a test with a high specificity (few false positive) has a low type I error rate. The number of genes selected refers to the genes identified as relevant by each method in the training set. The analysis of these genes gives information on prediction accuracy.

The Matthews correlation coefficient (MCC) is defined as

$$
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.
$$

The MCC measure is a global measure of accuracy, and a larger MCC indicates a better performance.

**Real Data Applications**

We applied the three network methods on different real datasets containing large-scale microarray gene expression measurements from ovarian and breast cancer including survival information (see Table 1) in order to facilitate the detection of molecular biomarker and pathway analysis with clinical utility.
TABLE 1 | Microarray Dataset Summary (OS = overall survival).

| Datasets   | Ref.                                      | Sample number | Platform                  | Genes number | Survival data | Cancer type |
|------------|-------------------------------------------|---------------|---------------------------|--------------|---------------|-------------|
| GSE26712   | Bonome et al., 2008                       | 185           | Affymetrix U133A          | 13104        | OS            | Ovarian     |
| OV-TCGA    | The Cancer Genome Atlas Research Network, 2011 | 578           | Affymetrix U133A          | 13104        | OS            | Ovarian     |
| GSE20685   | Kao et al., 2011                          | 327           | Affymetrix U133Plus2      | 21686        | OS            | Breast      |
| GSE7390    | Desmedt et al., 2007                      | 198           | Affymetrix U133A          | 13718        | OS            | Breast      |

Ovarian Datasets

We downloaded the first ovarian dataset from NCBI Gene Expression Omnibus as raw .CEL files (Bonome: GSE26712). The data contain gene expression profiling for extensive set of 185 primary ovarian tumors untreated late-stage (III–IV) high-grade (2,3) patients hospitalized at the Memorial Sloan-Kettering Cancer Center between 1990 and 2003. The Affymetrix human U133A microarray platform was used. The second ovarian dataset, the ovarian TCGA, was downloaded from The Cancer Genome Atlas data portal (The Cancer Genome Atlas Research Network, 2011). It was obtained at the gene level (level 3) using the Affymetrix human U133A microarray from 578 samples. All patients were diagnosed with high-grade serous carcinoma and were in an advanced stage. We noted that such datasets are very similar in terms of type of patients, platforms, and cancer disease. Therefore, they can be also used for validation.

Breast Datasets

The breast cancer microarray datasets were downloaded from NCBI GEO database as raw .CEL files (Kao: GSE20685 and Desmedt: GSE7390). Gene expression profiling of the first dataset was conducted on fresh frozen breast cancer tissue collected from 327 patients diagnosed and treated between 1991 and 2004 at the Koo Foundation Sun-Yat-Sen Cancer Center. Hybridization targets were prepared from total RNA according to the Affymetrix U133 plus 2.0 platform. The second breast cancer dataset was chosen on gene expression profiling of frozen samples from 198 N—systemically untreated patients at the Bordet Institute. It was based on the Affymetrix U133 platform.

Preprocessing

All the raw files were processed and normalized by RMA package available in Bioconductor (Gentleman et al., 2004). Between arrays normalization was carried out by using the preprocessCore package available in Bioconductor (Gentleman et al., 2004). Survival data (OS, i.e., overall survival), censoring indicator and time to death, for each patients in every dataset were also given (Figure 2, step 1).

Cancer Genes and Related Functional Networks

Following our previous study (Iuliano et al., 2014), in order to better analyze real datasets, we first applied a biologically inspired size reduction of the dataset, then we built an a-priori network information for the type of cancer under investigation (see Figure 2, step 1). For a better focus on genes that are more likely to be relevant in cancer, we selected the high-risk cancer genes using the Human Experimental/Functional Mapper (Huttenhower et al., 2009), which is based on a regularized Bayesian integration system. This mapper provides a p-value for each gene describing the significance of the relation between the gene and the disease of interest (breast and ovarian cancer, respectively). In our analysis, we selected only the genes with p < 0.05. A summary of the final number of the genes selected from each dataset is reported in Table 2. The network matrices used to test the network-based Cox models in our analysis were also derived from the Human Experimental/Functional Mapper which provides maps describing the genes functional activity and interaction networks in over 200 areas of human cellular biology with information from 30,000 genome-scale experiments. This functional network summarizes information from a variety of biologically informative perspectives: prediction of protein function and functional modules, cross-talk among biological processes, and association of novel genes and pathways with known genetic disorders (Huttenhower et al., 2009). The edges of the network are weighted between [0, 1] and express the functional relation between two genes. Note that the functional linkage network includes more information than Human PPI, frequently used as the network prior knowledge. It is clear that taking into account such biological knowledge helps in identifying significant genes that are functionally related in order to obtain important results biologically interpretable.

In order to adapt the gene network to the different methods, the final weight matrix was slightly different from method to method. In particular, since AdaLnet requires a weight matrix consisting of 0 and 1, each matrix element was set equal to 0 (or 1) if the weight value was below (or above) a fixed threshold equals to 0.5. On the other hand, Net-Cox uses the original weight matrix as obtained in the original paper (Huttenhower et al., 2009).

RESULTS

In our study, we analyzed three network-based Cox regression methods described in Section Methods both on simulated
TABLE 3 | Simulation results for Not-Overlapped settings. Sensitivity, specificity, number of selected genes, false positive rates, and MCC were averaged over the 100 replications.

| Method    | Sensitivity | Specificity | No. genes | No. FP      | MCC        |
|-----------|-------------|-------------|-----------|-------------|------------|
| Net-Cox   |             |             |           |             |            |
| No. genes = 44 | 0.240 (0.042) | 0.977 (0.004) | 44,000 (0.000) | 22,910 (3.677) | 0.300 (0.063) |
| No. genes = 88 | 0.489 (0.071) | 0.956 (0.006) | 88,000 (0.000) | 44,940 (6.233) | 0.445 (0.077) |
| No. genes = 176 | 0.737 (0.087) | 0.890 (0.008) | 176,000 (0.000) | 111,180 (7.692) | 0.464 (0.070) |
| AdaLnet   |             |             |           |             |            |
| General setting | 0.444 (0.250) | 0.792 (0.170) | 249,360 (193.786) | 210,330 (172.384) | 0.190 (0.059) |
| No. genes ≤ 100 | 0.200 (0.085) | 0.967 (0.021) | 51 (27.256) | 33,395 (21.227) | 0.220 (0.064) |
| No. genes > 100 | 0.627 (0.160) | 0.660 (0.099) | 399 (113.254) | 343,807 (100.118) | 0.166 (0.041) |
| fastcox   |             |             |           |             |            |
| General setting | 0.141 (0.117) | 0.970 (0.037) | 42.62 (46.613) | 30.19 (37.833) | 0.160 (0.082) |
| No. genes ≤ 10 | 0.017 (0.017) | 0.999 (0.0002) | 1.524 (1.486) | 0.048 (0.216) | 0.099 (0.07) |
| No. genes > 10 | 0.231 (0.063) | 0.949 (0.036) | 72,379 (40.331) | 52,017 (36.492) | 0.204 (0.054) |

The table reports three consensus rankings for Net-Cox obtained selecting 44, 88, and 176 genes. For AdaLnet and fastcox, we show the results related to the general setting, and the statistical measures obtained when the number of selected genes is higher (or lower) of a fixed threshold (threshold was set equal to 100 for AdaLnet and equal to 10 for fastcox). Standard deviation is reported in brackets.

TABLE 4 | Simulation results for overlapped settings.

| Method    | Sensitivity | Specificity | No. genes | No. FP      | MCC        |
|-----------|-------------|-------------|-----------|-------------|------------|
| Net-Cox   |             |             |           |             |            |
| No. genes = 44 | 0.156 (0.043) | 0.970 (0.004) | 44,000 (0.000) | 30,240 (3.766) | 0.175 (0.064) |
| No. genes = 88 | 0.288 (0.044) | 0.938 (0.004) | 88,000 (0.000) | 62,620 (3.842) | 0.227 (0.048) |
| No. genes = 176 | 0.386 (0.044) | 0.860 (0.003) | 176,000 (0.000) | 114,010 (3.860) | 0.182 (0.035) |
| AdaLnet   |             |             |           |             |            |
| General setting | 0.262 (0.178) | 0.879 (0.144) | 145,280 (160.666) | 122,240 (145.679) | 0.166 (0.067) |
| No. genes ≤ 100 | 0.141 (0.064) | 0.977 (0.020) | 35,635 (24.760) | 23,206 (20.296) | 0.196 (0.060) |
| No. genes > 100 | 0.467 (0.108) | 0.713 (0.105) | 331,973 (114.325) | 290,865 (106.135) | 0.114 (0.043) |
| fastcox   |             |             |           |             |            |
| General setting | 0.098 (0.099) | 0.974 (0.039) | 34,55 (47.732) | 25.89 (39.807) | 0.134 (0.061) |
| No. genes ≤ 10 | 0.019 (0.015) | 0.999 (0.0001) | 1.679 (1.281) | 0.017 (0.134) | 0.115 (0.065) |
| No. genes > 10 | 0.199 (0.061) | 0.942 (0.040) | 76,386 (45.224) | 58,818 (40.830) | 0.158 (0.044) |

Sensitivity, specificity, number of selected genes, false positive rates and MCC were averaged over the 100 replications. The table reports three consensus rankings for Net-Cox obtained selecting 44, 88, and 176 genes. For AdaLnet and fastcox, we show the results related to the general setting, and the statistical measures obtained when the number of selected genes is higher (or lower) that a fixed threshold (threshold was set equal to 100 for AdaLnet and equal to 10 for fastcox). Standard deviation is reported in brackets.

Simulation Studies

We analyze the performance of the three analyzed methods in two simulation settings where the number of relevant genes is fixed a-priori to 88 genes. The first setting simulates a scenario with not overlapped pathways, which means that each gene in the network belongs to only one pathway (not-overlapped pathways). The second setting represents a more realistic scenario with a set of genes shared among different pathways (overlapped pathways). In both cases, a five-fold cross validation was conducted on the full dataset in order to select the tuning parameters ($\lambda$, $\alpha$) and to obtain the coefficient estimates to biomedical community in the form of a comprehensive methodological procedure (see Figure 1).
by using the three methods. The details of the simulation data are reported in Section Methods.

The performance of each method is summarized by several statistical measures: sensitivity, specificity, number of selected genes, false positive rates, and Matthews correlation coefficient (MCC). Simulation results for both the models are reported in Tables 3, 4, respectively (standard deviation is reported in brackets). To analyze the signature genes identified by Net-Cox, which is a method based on ridge regression, we considered three different consensus rankings where the number of significant genes selected by the method was fixed to 44, 88, and 176 genes, respectively. The selected genes were classified in descending order according to the absolute value of the regression coefficients. On the other hand, to better highlight the variable selection performance of AdaLnet and fastcox, we split the 100 iterations in two groups based on the number of genes selected at each iteration. We fixed 100 genes as threshold for AdaLnet and 10 genes for fastcox, then we computed again the statistical measures based on the two groups.

In the not-overlapped setting, Net-Cox performed better than the other two methods as showed by the MCC, which provides an overall measure of accuracy. In particular, when considering 44 and 88 genes, the false positive rate in Net-Cox was 22.910 and 44.940, respectively, with MCC equals to 0.300 and 0.445. Sensitivity and specificity were, respectively, 0.240 and 0.977 in the first case, 0.489 and 0.956 in the second case study. When the number of selected genes was increased to 176, even if the false positive rate increased resulting in a lower specificity (0.890), the sensitivity reached its highest values producing the highest MCC (0.464).

Since the majority of the selected genes were irrelevant and both AdaLnet and fastcox resulted in sparse models, specificity was much higher than sensitivity and was comparable between the two variable selection methods. In particular, in the not-overlapped setting, AdaLnet selected in average 249.360 genes with a false positive rate equals to 210.330. Sensitivity and specificity were equal to 0.444 and 0.792 resulting in a MCC of 0.190. On the other hand, fastcox selected in average 42.62 genes with a false positive rate of 30.19. MCC was equal to 0.160 with sensitivity 0.141 and specificity 0.970.

AdaLnet had the best performance when the number of selected genes was below 100, while fastcox exhibit the best performance when the number of genes was above 10. This means that in the other cases the methods fail in the execution of the cross-validation (see Supplementary Image 1).

In the overlapped-pathways setting, Net-Cox obtained the highest MCC overall when considering 88 genes (MCC equals to 0.227) with a false positive rate equals to 62.620, sensitivity 0.288 and specificity 0.938. However, even if the specificity levels of the three consensus rankings were almost equal to the previous setting (specificity for 44, 88, and 176 genes equals to 0.970, 0938, and 0.860, respectively), in this setting Net-Cox sensitivity decreased resulting in lower MCC compared to the not-overlapped case (MCC for 44, 88, and 176 genes equals to 0.175, 0.227, and 0.182, respectively). AdaLnet and fastcox also reported lower MCCs compared to the not-overlapped setting (MCC equals to 0.166 in AdaLnet and 0.134 in fastcox). In particular, both AdaLnet and fastcox showed an higher specificity than before (0.879 and 0.974, respectively) but a lower sensitivity (0.262 and 0.098). Further analysis showed that AdaLnet had the highest MCC when the number of selected genes was below 100 (MCC 0.196), while fastcox had the highest MCC (0.158) when the number of selected genes was above 10, in accordance with the previous results (see Supplementary Image 2).

### Real Data Analysis

In order to evaluate the performance of the three Cox models in terms of survival analysis, we used cross-validated Kaplan–Meier curves (Simon et al., 2011b) for overall survival (OS) both on ovarian and breast microarray studies (see Figure 2, step 2). Note that p-value was estimated within the same dataset but the crossvalidation approach is used to correct over optimistic conclusions due to the lack of independence between samples.

Moreover, since the ovarian datasets are comparable in terms of types of patients, platforms and cancer disease, Kaplan–Meier curves and two-side log-rank test were used to estimate the survival time and stratify the low-risk and high-risk groups on the independent test set (see Figure 2, step 3).

Table 5 reports the number of genes selected by the three Cox regression methods for each OS and the optimal tuning parameter $\alpha$. Interestingly, the optimal $\alpha$ was often equal to 0.5, indicating that there was a good balance between statistical constraints and network information. These results confirm that the network carries important information useful for improving survival analysis. Moreover, since Net-Cox is a method based on ridge regression, the genes are only shrunk and it is necessary to fix a threshold for selecting the most relevant cancer genes. Hence, within each fold, we ordered the genes according to the absolute value of the corresponding regression coefficients, then we considered the union of the top 50 genes selected in each fold.

### Table 5 | Optimal $\alpha$ cross-validated value calculated on the k training sets.

| Datasets | $k$ Partitions | Net-Cox | AdaLnet | fastcox |
|----------|----------------|---------|---------|---------|
|          | $\alpha$ | Genes selected | $\alpha$ | Genes selected | $\alpha$ | Genes selected |
| GSE26712 | 5 | 0.2 | 101 | 0.5 | 23 | 0.01 | 453 |
| OV-TCGA  | 5 | 0.5 | 99 | 0.5 | 38 | 0.1 | 623 |
| GSE206865 | 5 | 0.5 | 76 | 0.5 | 28 | 0.01 | 298 |
| GSE7390  | 5 | 0.5 | 89 | 0.5 | 14 | 0.01 | 423 |
Iuliano et al. Network-Based Cox Regression

FIGURE 3 | Cross-validated Kaplan–Meier curves of the prognostic models on GSE26712 dataset. The patients are divided in high-risk and low-risk groups based on the pathways and genes selected by each method for overall survival (OS). The survival probabilities of these two groups are compared using the log-rank test by using Net-Cox (A), AdaLnet (B), and fastcox (C).

In the following, we present the main results obtained.

Results on the Ovarian Datasets

Figures 3, 4 show the cross-validated Kaplan–Meier curves for high-and-low risk groups patients selected in the ovarian datasets (Benome: GSE26712 and OV TCGA datasets, respectively).

Figure 3 shows that in the Bonome dataset the gap between the survival curves of the two risk groups in Net-Cox (Figure 3A) and fastcox (Figure 3C) is wider compared to AdaLnet (Figure 3B). In particular, in predicting survival probabilities, fastcox (permuted $p < 0.05$) seem to discriminate the risk groups better than Net-Cox and AdaLnet where the permuted $p > 0.05$. These
findings confirm the results previously obtained in Iuliano et al. (2014), in relation to the survival curves for each method. This was mainly due to the cross-validation approach used in this analysis to overcome the sample splitting problem with too small dataset.

On the other hand, in the OV TCGA dataset (Figure 4), the survival curves for high-and-low risk patients are not significantly separated. In particular, fastcox is the only method with a significant difference (permuted $p < 0.05$) in the OS between the high-and-low-risk groups.

Finally, to test the survival prediction across independent datasets, we used the ovarian OV TCGA dataset as training set, and the Benome dataset as the test set to predict the risk scores.
Results on Breast Datasets

Figures 6, 7 show the cross-validated Kaplan–Meier curves for high-and-low risk groups patients selected in the breast datasets (Kao: GSE20685 and Desmedt: GSE7390, respectively). In the Kao dataset, the permuted p-value related to Figure 6A (Net-Cox) and Figure 6C (fastcox) was smaller than 0.05, which means the high-risk and low-risk groups were significantly separated and the selected pathways and genes were related to survival times. In Figure 6B (AdaLnet), a patient of the high-risk group fell in the low-risk group and the permuted p-value is not significant.

We performed the same analysis for high-and-low risk patients in the Desmedt dataset. Also in this case, there was a significant difference in OS between the two risk groups as shown in Figure 7A (Net-Cox) and Figure 7C (fastcox) where the permuted p-value is smaller than 0.05. In Figure 7B (AdaLnet) the permuted p-value is not significant.

Identified Pathways

In this section, we present the results of the analysis in terms of KEGG pathways analysis based only on not-isolated genes (see section Methods for details). We report here only the networks related to AdaLnet and Net-Cox since all the networks related to fastcox have more than 100 node and 2000 edges and a clear visualization would not be possible. However, the lists of the genes selected by fastcox and the related pathways are reported in Supplementary Table 1 (ovarian datasets) and Supplementary Table 2 (breast datasets).

Figures 8, 9 show the gene-networks obtained for the Bonome dataset (GSE26712) built on the genes identified by Net-Cox and Adalnet, respectively. From the color of the nodes, we can infer that all the selected genes have a significant relation with ovarian cancer. Indeed, almost all the genes are close to red except for AKT3 which has a p-value correlation equal to 0.039. Indeed, AKT3 is usually involved in prostate and breast cancer (Nakatani et al., 1999). However, since it was selected both by Net-Cox and fastcox, a possible significant relation between AKT3 and ovarian cancer could be inferred as indeed confirmed by literature (Liby et al., 2012). In particular, AKT3 has a specific role in the genesis of ovarian cancer through modulation of G2-M phase transition (Cristiano et al., 2006).

As shown in Figure 8, AKT3 is also involved in many cancer pathways, such as KEGG basal cell carcinoma, KEGG prostate cancer, and KEGG melangiogenesis. It is worthy to note that this gene was also selected in our previous study (Iuliano et al., 2014) by all the analyzed methods and it was also involved in the same cancer related pathways. These findings confirm the importance of AKT3 in ovarian cancer as confirmed indeed by literature (Cristiano et al., 2006).

In the Bonome dataset (GSE26712), Adalnet selected only two not-isolated genes (RB1 and BRCA2) involved in two different cancer pathways (Figure 9). Both the genes have been frequently observed in epithelial ovarian cancer (Flesken-Nikitin et al., 2003; Dinulescu et al., 2005; Naora and Montell, 2005) and several studies report their stable correlation (Flesken-Nikitin et al., 2003; The Cancer Genome Atlas Research Network, 2011). Moreover, the strong interaction between RB1 and the tumor protein TP53 (Dong et al., 1997; Schuijer and Berns, 2003) has been identified by Net-Cox and fastcox (Figure 8).
FIGURE 8 | Gene-network of not isolated genes selected by Net-Cox in the Bonome ovarian dataset (GSE26712). Each node represents a gene and an edge between two nodes means that the two genes belong to the same pathway. Different colors are used for different pathways. The color of each node represents the $p$-value of the interaction between the gene and ovarian cancer (Huttenhower et al., 2009). Genes with $p > 0.10$ are represented in green.

FIGURE 9 | Gene-network of not isolated genes selected by Adalnet in the Bonome ovarian dataset (GSE26712). Each node represents a gene and an edge between two nodes means that the two genes belong to the same pathway. Different colors are used for different pathways. The color of each node represents the $p$-value of the interaction between the gene and ovarian cancer (Huttenhower et al., 2009). Genes with $p > 0.10$ are represented in green.

Figures 10, 11 show the gene-networks obtained for the OV TCGA ovarian dataset built on the genes identified by Net-Cox and Adalnet, respectively. As already observed in the Bonome dataset analysis, all the selected genes in the OV TCGA dataset resulted strongly correlated with ovarian cancer. Indeed, almost all the genes are close to red. The only gene with a slightly different color is FZD3 which has a $p$-value of 0.049 and was selected by all the three methods. Hence, even if this gene has been mainly classified as gastric-cancer-related (Katoh, 2005), our results prove that it also has a relevant effect in ovarian cancer as confirmed by literature (Tapper et al., 2001). It is also important to note that other genes have been selected by all the three methods (i.e., GMPR, ENPP1, and APC). Such genes have been already classified as ovarian-related in cancer literature (Gayther et al., 1997; Kikuchi et al., 2007; Rikova et al., 2007), but, in our analysis, the pathways involved in such relation are also investigated. For example, while GMPR and ENPP1 interact simply through the KEGG purine metabolism pathway, the APC-FZD3 interaction involves three different pathways: KEGG basal carcinoma, KEGG pathways in cancer, and KEGG wnt signaling pathway.

It is worthy to note that some of the genes selected by the three methods (e.g., NPY, COL5A1, EGFR, and FBL1) have been already reported in literature (Zhang et al., 2013) where an analysis of subnetwork signatures in ovarian cancer based on Cox model is presented. Moreover, our approach selected new genes, such as AKT3 and RB1, which are also related to ovarian cancer (Flesken-Nikitin et al., 2003; Cristiano et al., 2006). These results show that our findings are consistent with the previous ones including, at the same time, other gene signatures.

Figures 12, 13 report the gene-networks selected in the Kao dataset (GSE20685) by Net-Cox and Adalnet, respectively. FGFR2 and BCL2 were again selected in this dataset confirming the strong relevance of the two genes in breast cancer. Moreover,
FIGURE 10 | Gene-network of not isolated genes selected by Net-Cox in the TCGA ovarian dataset. Each node represents a gene and an edge between two nodes means that the two genes belongs to the same pathway. Different colors are used for different pathways. The color of each node represents the p-value of the interaction between the gene and ovarian cancer (Huttenhower et al., 2009). Genes with p > 0.10 are represented in green.

FIGURE 11 | Gene-network of not isolated genes selected by Adalnet in the TCGA ovarian dataset. Each node represents a gene and an edge between two nodes means that the two genes belong to the same pathway. Different colors are used for different pathways. The color of each node represents the p-value of the interaction between the gene and ovarian cancer (Huttenhower et al., 2009). Genes with p > 0.10 are represented in green.

BRCA2 (Wooster et al., 1995) was selected by Net-Cox and fastcox confirming the accuracy of our analysis. It is also worthy to note that in all the breast cancer gene-networks the KEGG prostate cancer is always recurrent. This is mainly due to the common biomarkers between the two diseases (Yang et al., 1998; Mattie et al., 2006) and through our analysis new common biomarkers can be identified.

In the Desmedt dataset (GSE7390), all the genes selected by Adalnet were isolated and no network was built in this case. A list of the genes selected is reported in Table 6.
reports the gene-network related to the genes selected by Net-Cox. All the selected genes show a strong relation with the disease, such as FGFR2 and BCL2, which were selected by both Net-Cox and fastcox and are involved in KEGG prostate cancer and in KEGG pathways in cancer. Both the genes are largely known as independent prognostic marker in breast cancer (Hunter et al., 2007; Thomadaki et al., 2007; Callagy et al., 2008). Both Net-Cox and fastcox selected UGT2B15, which has a breast-cancer-correlation $p = 0.049$. This gene has been usually involved in prostate cancer (Gsur et al., 2002), but recent works highlight its role also in breast cancer (Wegman et al., 2007).

In the analysis of the breast datasets, there was no overlap with our previous study (Iuliano et al., 2014). This was mainly due to the different datasets analyzed here potentially (different cancer subtype and different types of conditions) and to the more sophisticated procedures followed in this analysis. Indeed, in our previous work, we split the dataset in training and test set only
FIGURE 14 | Gene-network of not isolated genes selected by Net-Cox in the GSE7390 breast dataset. Each node represents a gene and an edge between two nodes means that the two genes belong to the same pathway. Different colors are used for different pathways. The color of each node represents the p-value of the interaction between the gene and breast cancer (Huttenhower et al., 2009). Genes with $p > 0.10$ are represented in green.

TABLE 6 | List of genes selected by Adalnet in the breast dataset GSE7390.

| Genes    | p-values |
|----------|----------|
| BRCA1    | 0        |
| GYPB     | 0.0489   |
| MYBL2    | 0.0026   |
| ADH6     | 0.0259   |
| GHRHR    | 0.0007   |
| GUCY2C   | 0.0323   |
| PPP2R1B  | 0.0321   |
| SLC1A2   | 0.0450   |
| SLC12A3  | 0.0483   |
| LIPF     | 0.0449   |
| TRIP13   | 0.0001   |
| PPM1E    | 0.0026   |
| CEP152   | 0.0064   |
| PSPC1    | 0.0475   |

The second column reports the breast-cancer correlation p-value of each gene accordingly (Huttenhower et al., 2009). All the selected genes resulted isolated and no network was built in this case.

once, while here we used a cross-validation procedure that is expected more robust results.

**DISCUSSION AND CONCLUSIONS**

A key issue in cancer survival analysis is uncovering the relation between gene expression profiles and cancer patients survival in order to identify biomarkers for disease diagnosis and treatment. In the last years, there has been a growing interest in methods that incorporate network information into classification algorithms for genes signature discovery. The main aims are to identify molecular biomarkers that reliably predict patient’s response to therapy and to avoid ineffective treatment for reducing drug side-effects and associated costs. For this purpose, prognostic and diagnostic biomarker signatures need to be derived from omics data for various disease entities in order to offer useful methodological and practical strategy in research and clinical settings.

Here, we presented an extended methodological strategy for the analysis of gene signatures and survival prediction (see Figure 1). We integrated a new cross-validation method (Simon et al., 2011b) with the most recent network penalized Cox models (Yang and Zou, 2012; Zhang et al., 2013; Sun et al., 2014) to obtain an effective multi-splitting of the data and achieve an accurate survival prediction (see Figure 2). The analysis of the models was based both on simulated and real datasets in order to provide an accurate analysis in terms of statistical and biological investigation. Indeed, we showed that, given a number of variables not extremely high, all the analyzed methods were able to select the altered genes under different simulation settings. On the other hand, the analysis on real cancer datasets showed that through the integration of network information into Cox regression methods it is possible to identify cancer gene signatures with an accurate prognostic performance. Therefore, the contribution of this study is two-fold. Firstly, to obtain an integrative analysis of cancer genes networks and survival prediction. Secondly, to provide a computational and methodological framework for better investigating cancers regulatory networks and facilitating the management of patients in terms of prognosis, diagnosis and treatment.

The findings of this study have a number of important implications for future practice. Firstly, a practically appealing study based on a fast screening procedure (Fan and Lv, 2008; Fan et al., 2010) could be introduced in order to reduce the size of the feature space to a moderate scale. In fact, several types of screening procedures could be combined to integrate biological information into statistical screening analysis and provide more definitive understanding of the gene-regulatory networks. Secondly, the integration of clinical information and data from different omics (e.g., epigenomics or metabolomics) into the screening procedure could also provide a more accurate investigation and prevent the drawbacks of the current methods. Moreover, a more accurate biomarkers investigation could be performed using a number of high-quality
binary PPIs available in literature (Rolland et al., 2014) where a proteome-scale map of the human binary interactome is compared to alternative network maps in order to give a deeper insight into genotype-phenotype relationships. Finally, it will be necessary to develop an user-friendly interface to turn this methodological framework into a practical tool.

**AUTHOR CONTRIBUTIONS**

AI and AO are joint first authors and both authors contributed equally. AI and AO prepared the computational codes and carried out all of the statistical analysis. CA, ID, and PL initiated and coordinated the work, guided the study design, supervised all data curation and analysis, and finalized all study conclusion. CA, ID, and PL are equal contributors. All the authors wrote, reviewed and approved the final manuscript.

**REFERENCES**

Bonome, T., Levine, D. A., Shih, J., Randonovich, M., Pise-Masison, C. A., Bogomolniy, F., et al. (2008). A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res. 68, 5478–5486. doi: 10.1158/0008-5472.CAN-07-6595

Callagy, G. M., Webber, M. J., Pharoah, P. D., and Caldas, C. (2008). Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer. BMC Cancer 8:153. doi: 10.1186/1471-2407-8-153

Candes, E., and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. 35, 2313–2351. doi: 10.1214/0090536060001523

Cox, D. R. (1972). Regression models and life-tables. J. R. Stat. Soc. B Methodol. 187–220.

Cristiano, B. E., Chan, J. C., Hannan, K. M., Lundie, N. A., Marmy-Conus, N. J., Campbell, I. G., et al. (2006). A specific role for AKT3 in the genesis of ovarian cancer through modulation of G2-M phase transition. Cancer Res. 66, 11718–11725. doi: 10.1158/0008-5472.CAN-06-1968

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemant, F., Haibe-Kains, B., et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. Clin. Cancer Res. 13, 3207–3214. doi: 10.1158/1078-0432.CCR-06-2765

Dinulescu, D. M., Ince, T. A., Quade, B. J., Shafer, S. A., Crowley, D., and Jacks, T. (2005). Role of K-ras and pten in the development of mouse models of endometriosis and endometrioid ovarian cancer. Nat. Med. 11, 63–70. doi: 10.1038/nm1173

Dong, Y., Walsh, M. D., McGuckin, M. A., Cummings, M. C., Gabrielli, B. G., Wright, G. R., et al. (1997). Reduced expression of retinoblastoma gene product (pRb) and high expression of p53 are associated with poor prognosis in ovarian cancer. Int. J. Cancer, 74, 407–415. doi: 10.1002/(SICI)1097-0215(19970822)74:4<407::AID-IJC8>3.0.CO;2-Z

Engler, D., and Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. Stat. Appl. Genet. Mol. Biol. 8, 1–22. doi: 10.2202/1544-6151.1423

Fan, J., Feng, Y., and Wu, Y. (2010). "High-dimensional variable selection for cox's proportional hazards model," in Borrowing Strength: Theory Powering Applications-A Festschrift for Lawrence D. Brown, eds J. O. Berger, T. T. Cai, and I. M. Johnstone (Beachwood, OH: Institute of Mathematical Statistics), 70–86.

Fan, J., and Li, R. (2001). Variable selection via concave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360. doi: 10.1198/016214501753382273

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. B Stat. Methodol. 70, 849–911. doi: 10.1111/j.1467-9868.2008.00674.x

Flesken-Nikitin, A., Choi, K.-C., Eng, J. P., Shmidt, E. N., and Nikitin, A. Y. (2003). Induction of carcinogenesis by concurrent inactivation of p53 and Rb1 in the mouse ovarian surface epithelium. Cancer Res. 63, 3459–3463.

Frohlich, H. (2014). Including network knowledge into Cox regression models for biomarker signature discovery. Biom. J. 56, 287–306. doi: 10.1002/bimj.201300035

Gayther, S. A., Mangion, J., Russell, P., Seal, S., Barfoot, R., Ponder, B. A., et al. (1997). Variation of risks of breast and ovarian cancer associated with different germline mutations of the BRCA2 gene. Nat. Genet. 15, 103–105. doi: 10.1038/ng0197-103

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5:R80. doi: 10.1186/gb-2004-5-10-r80

Gong, H., Wu, T. T., and Clarke, E. M. (2014). Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression. BMC Syst. Biol. 8(Suppl. 1):S3. doi: 10.1186/1752-0509-8-s1-s3

Gir, A., Preyer, M., Haidinger, G., Schatzl, G., Madersbacher, S., Marberger, M., et al. (2002). A polymorphism in the UDP-glucuronosyltransferase 2B15 gene (DGUS2) is not associated with prostate cancer risk. Cancer Epidemiol. Biomarkers Prev. 11, 497–498.

Gui, J., and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21, 3001–3008. doi: 10.1093/bioinformatics/bti422

Hudis, C. A. (2007). Trastuzumab mechanism of action and use in clinical practice. N. Engl. J. Med. 357, 39–51. doi: 10.1056/NEJMra043186

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. 39, 870–874. doi: 10.1038/ng2075

Huttenhower, C., Haley, E. M., Hibbs, M. A., Dumeaux, V., Barrett, D. R., Coller, H. A., et al. (2009). Exploring the human genome with functional maps. Genome Res. 19, 1093–1106. doi: 10.1101/gr.082214.108

Iuliano, A., Occipinti, A., Angelini, C., De Feis, I., and Liò, P. (2014). “Applications of network-based survival analysis methods for pathways detection in cancer,” in Computational Intelligence Methods for Bioinformatics and Biostatistics, eds C. Di Serio, P. Liò, A. Nonis, and R. Tagliaferri (Springer), 76–88.

**ACKNOWLEDGMENTS**

This research was partially supported by BioforIU Project and by EPIGEN Project. AO acknowledges the Qualcomm Research Scholarship and Cambridge Home and European Scholarship Scheme (CHESS). AI and AO would like to thank Prof. Pietro Liò for the hospitality at the Computer Laboratory in Cambridge and CA and ID for the hospitality at the IAC in Naples. We would like to thank also Prof. Hokeun Sun for sharing AdaLnet code and the reviewers for their insightful comments. PL received funding from the EU FP7-Health-F5-2012 under grant no. 305280 (MIMOmics).

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fphys.2016.00208
Iuliano et al.  

Network-Based Cox Regression

Jeong, H.-H., Kim, S. Y., Wee, K., and Sohn, K.-A. (2015). Investigating the utility of clinical outcome-guided mutual information network in network-based Cox regression. BMC Syst. Biol. 9:1. doi: 10.1186/1752-0509-9-S1-P8

Kao, K.-J., Chang, K.-M., Hsu, H.-C., and Huang, A. T. (2011). Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 11:143. doi: 10.1186/1471-2407-11-143

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., O’Callaghan, C. J., Tu, D., Tebbutt, N. C., et al. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N. Engl. J. Med. 359, 1757–1765. doi: 10.1056/NEJMoa0804385

Katoh, M. (2005). WNT/PCP signaling pathway and human cancer (review). Oncol. Rep. 14, 1583–1588. doi: 10.3892/or.14.6.1583

Kearns, M., and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural Comput. 11, 1427–1435. doi: 10.1162/089976699300016304

Kikuchi, R., Tsuda, H., Kanai, Y., Kasamatsu, T., Sengoku, K., Hirohashi, S., et al. (2005). Promoter hypermethylation contributes to frequent inactivation of a putative conditional tumor suppressor gene connecting tissue growth factor in ovarian cancer. Cancer Res. 67, 7095–7105. doi: 10.1158/0008-5472.CAN-06-4567

Kohavi, R. (1995). “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in IJCAI, Vol. 14 (Stanford, CA), 1137–1145.

Li, C., and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24, 1175–1182. doi: 10.1093/bioinformatics/btn081

Li, C., and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. Ann. Appl. Stat. 4, 1498. doi: 10.1214/10-AOAS332

Liby, T. A., Syrporopoulos, P., Buff Lindner, H., Eldridge, J., Beeson, C., Hsu, T., et al. (2012). Akt3 controls vascular endothelial growth factor secretion and angiogenesis in ovarian cancer cells. Int. J. Cancer 130, 532–543. doi: 10.1002/ijc.26010

Martinez-Ledesma, E., Verhaak, R. G., and Treviño, V. (2015). Identification of a multi-cancer expression biomarker for cancer clinical outcomes using a network-based algorithm. Sci. Rep. 5:11966. doi: 10.1038/srep11966

Mattie, M. D., Benz, C. C., Bowers, J., Sensinger, K., Wong, L., Scott, G. K., et al. (2006). Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. Mol. Cancer 5:24. doi: 10.1186/1476-4598-5-24

Molina, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301–3307. doi: 10.1093/bioinformatics/bti499

Mukhopadhyay, K., Thompson, D. A., Barthel, A., Sakaue, H., Liu, W., Weigel, R. J., et al. (1999). Up-regulation of Akt3 in estrogen receptor-deficient breast cancers and androgen-independent prostate cancer lines. J. Biol. Chem. 274, 21528–21532. doi: 10.1074/jbc.274.31.21528

Naora, H., and Montell, D. J. (2005). Ovarian cancer metastasis: integrating insights from disparate model organisms. Nat. Rev. Cancer 5, 355–366. doi: 10.1038/nrc1611

Raghubhadii, W., and Raghubhadii, V. (2014). Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. 2:3. doi: 10.1186/2047-2501-2-3

Refaeli Zadeh, P., Tang, L., and Liu, H. (2009). “Cross-validation,” in Encyclopedia of Database Systems, eds L. Liu and M. T. Ozsu (New York, NY: Springer), 532–538. doi: 10.1002/9780470037996.0-9.565

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., et al. (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. J. Biol. Chem. 282, 1873–1880. doi: 10.1074/jbc.M606150200

Simon, R., Friedman, J., Hastie, T., and Tibshirani, R., (2011a). Regularization using a network-based algorithm. J. stat. Plan. Pred. 131, 1190–1203. doi: 10.1016/j.jcpb.2014.10.050

Copyright © 2016 Iuliano, Ochepinti, Angelini, De Feis and Liò. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.