frontiers in Genetics

METHODS
published: 14 June 2018
doi: 10.3389/fgene.2018.00206

Combining Pathway Identification and Breast Cancer Survival Prediction via Screening-Network Methods

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

[1] Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche, Naples, Italy, [2] Telethon Institute of Genetics and Medicine, Pozzuoli, Italy, [3] Computer Laboratory, University of Cambridge, Cambridge, United Kingdom

Breast cancer is one of the most common invasive tumors causing high mortality among women. It is characterized by high heterogeneity regarding its biological and clinical characteristics. Several high-throughput assays have been used to collect genome-wide information for many patients in large collaborative studies. This knowledge has improved our understanding of its biology and led to new methods of diagnosing and treating the disease. In particular, system biology has become a valid approach to obtain better insights into breast cancer biological mechanisms. A crucial component of current research lies in identifying novel biomarkers that can be predictive for breast cancer patient prognosis on the basis of the molecular signature of the tumor sample. However, the high dimension and low sample size of data greatly increase the difficulty of cancer survival analysis demanding for the development of *ad-hoc* statistical methods. In this work, we propose novel screening-network methods that predict patient survival outcome by screening key survival-related genes and we assess the capability of the proposed approaches using METABRIC dataset. In particular, we first identify a subset of genes by using variable screening techniques on gene expression data. Then, we perform Cox regression analysis by incorporating network information associated with the selected subset of genes. The novelty of this work consists in the improved prediction of survival responses due to the different types of screenings (i.e., a biomedical-driven, data-driven and a combination of the two) before building the network-penalized model. Indeed, the combination of the two screening approaches allows us to use the available biological knowledge on breast cancer and complement it with additional information emerging from the data used for the analysis. Moreover, we also illustrate how to extend the proposed approaches to integrate an additional omic layer, such as copy number aberrations, and we show that such strategies can further improve our prediction capabilities. In conclusion, our approaches allow to discriminate patients in high-and low-risk groups using few potential biomarkers and therefore, can help clinicians to provide more precise prognoses and to facilitate the subsequent clinical management of patients at risk of disease.

Keywords: breast cancer, cox regression, high-dimensionality, network-penalized methods, screening techniques, survival analysis, pathway analysis

OPEN ACCESS

Edited by:
Wencong Chen,
Baylor Scott and White Research Institute (BSWRI), United States

Reviewed by:
Marco Scutari,
University of Oxford, United Kingdom
Alexey Goltsov,
Abertay University, United Kingdom

***Correspondence:**
Antonella Iuliano
a.iuliano@na.iac.cnr.it

†These authors have contributed equally to this work and are joint first authors.

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *05 March 2018*
Accepted: *24 May 2018*
Published: *14 June 2018*

Citation:
Iuliano A, Occhipinti A, Angelini C, De Feis I and Liò P (2018) Combining Pathway Identification and Breast Cancer Survival Prediction via Screening-Network Methods. Front. Genet. 9:206. doi: 10.3389/fgene.2018.00206

1. INTRODUCTION

Understanding the multidimensional complexity of breast cancer is an ongoing pursuit for many researchers to model survival oncological data. Technology advances offer great opportunities to explain cancer mechanisms, although there are significant challenges in extracting knowledge from such massive data and evaluating the findings. In the last years, a huge amount of genome-wide data collected using a variety of high-throughput technologies has been made publically available thanks to the effort of several international projects and consortia. For example, The Cancer Genome Atlas (TCGA) (Network, 2011, 2012, 2013), the Catalog of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2010), The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012) and many others projects were established to profile large tumor sets for different omics layers, such as gene expression, DNA structure and methylation, etc. By using these types of data, biological interaction networks based on physical interactions, such as protein-protein interactions, protein-DNA interactions, and phosphorylation can be also constructed. In particular, functional interaction networks connect genes with similar or related functions and are typically inferred from multiple sources, including co-expression, KEGG pathways, functional linkage, Gene Ontology (GO) terms, etc. Overall, data from these databases not only allow to better understand the deregulation of cellular mechanisms during diseases progression, but also provide opportunities for developing novel statistical and computational algorithms for the analysis of patient omic data and for the interpretation and validation of results.

Despite this progress, many cancer diseases do not have effective treatments yet. Recently, precision medicine has been used by clinicians to take all kinds of decisions regarding the patient management and therapeutic treatments (Huang et al., 2016). In particular, prognostic biomarkers have been used for more effective selection of patient subgroups with different therapeutic strategies. In the recent past, inference was carried out by looking at a specific omic type, such as gene expression or DNA structural variations, etc. Nowadays, it is clear that multi-omic data integration is becoming necessary to investigate the genomic mechanisms involved in complex diseases (Angelini and Costa, 2014). From a statistical point of view, one of the most important challenges in integrating multi-omic profiles is to cope with the high-dimensionality of the data. To overcome this issue and optimize model predictions, innovative statistical approaches have been developed (Pineda et al., 2015). Indeed, taking more levels into account increases the dimensionality of the problem and requires additional steps for data compatibility, normalization, correction and integration (Ritchie et al., 2015; Bersanelli et al., 2016).

To reduce dimensionality from a high to a moderate scale, one can use feature screening by ranking the significant genes based on their marginal associations with the outcome variable and removing unimportant genes from the bottom of an ordered list. Feature screening techniques in ultrahigh-dimensional data analysis were introduced in Fan and Lv (2008), where the sure independence screening (SIS) and the SIS screening were proposed when the data come from an ordinary linear model with normal errors. Then, such techniques were extended to generalized linear models (Fan et al., 2009, 2010b). Nonparametric independence screening in sparse ultrahigh-dimensional additive models was presented in Fan et al. (2011). In that article, the authors suggested estimating the nonparametric components marginally with spline approximation, and ranking the importance of predictors using the magnitude of nonparametric components. A sure independent ranking and screening (SIRS) procedure to screen significant predictors in multi-index models was proposed in Zhu et al. (2011). The authors showed that under the assumption of linearity on the predictor vector, the SIRS satisfies the ranking consistency property. Finally, a sure screening procedure for Cox's proportional hazards model was presented in Fan et al. (2010a), Zhao and Li (2012) and Song et al. (2014) in order to understand the association between genomic information and survival information on oncological patients. In this work, we present three screening inspired approaches that turn out to be useful in reducing the dimensionality of the data.

Nevertheless, when the number of variables (i.e., genes or genomic features) p is much larger than the observations (i.e., patients) n ($p \gg n$), the Cox model (Cox, 1972) cannot be applied directly. Therefore, alternative methods combining penalized Cox regression models and variable selection have been developed. Those methods include ℓ_1 and ℓ_2 norms (Zou and Hastie, 2005; Simon et al., 2011; Wu, 2012), the SCAD (Fan and Li, 2001), the adaptive Lasso (Zou, 2006) and the Dantzig selector (Candes and Tao, 2007) which have been proposed to infer parameters in order to further reduce the feature space and to impose sparsity on the solutions. Such penalized approaches improve prediction capabilities and interpretability of results when a large number of variables is present. Similar approaches might also be used when there exist an underlying structure on the gene/feature space, for example to account for gene regulatory mechanisms or patterns of co-expressions. In this framework, the correlation structure can be specified as constraints to the Cox model (Zhang et al., 2013; Fröhlich, 2014; Gong et al., 2014; Sun et al., 2014). Therefore, the regulatory genomic information is encoded by a network, where genes are depicted as nodes and their pair-wise relations as edges connecting two nodes. The network is converted in a Laplacian matrix and is used as penalty in the Cox regression models. In particular, the network can represent different types of relationships such as gene expression correlations, KEGG pathways information, functional interaction networks or Protein-Protein Interaction networks. Generally, the Cox regression models built on biological networks are called "network-based Cox regression models". For instance, a comprehensive overview of computational methods used for biomarkers identification, including rank-based feature selection methods and major network methodologies used in system biology was performed in Guo and Wan (2014).

In this article, we combine screening techniques and network-penalized approaches for building novel methods that select subsets of genes associated to patients survival in cancer (see **Figure 1**). In particular, we use METABRIC training set containing a long-term follow-up of about 1,000 breast cancer

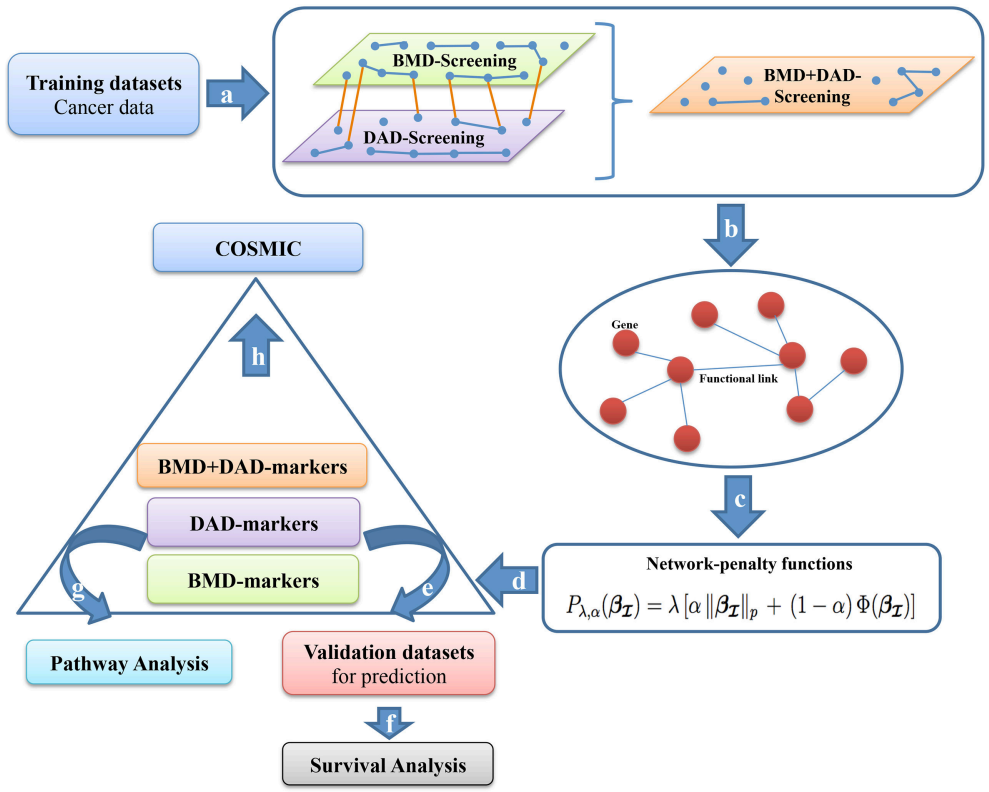

FIGURE 1 | Screening-network procedure. (a) BMD-screening, DAD-screening or BMD+DAD-screening are applied to the training dataset T to choose a subset of genes $\{x_j, j \in \mathcal{I}\}$. (b) We use subset $\{x_j, j \in \mathcal{I}\}$ to incorporate the biological information as networks of interactions. (c) Network-penalized methods are executed on $\{x_j, j \in \mathcal{I}\}$. (d) High-risk genes or potential biomarkers ($\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$) are selected (from each screening approach) and are used to separate patients in two groups (high-and-low risk group). (e–f) The performance of the survival prediction are assessed using a testing dataset D. To make sense of the gene signatures (g) Pathway analysis is performed on high-risk genes and (h) COSMIC investigation is carried out.

patients (Curtis et al., 2012), after having applied different types of screenings, we fit a network-penalized model for identifying gene signatures predicting survival responses. Then, we validate the capabilities of the proposed methods using about other 1,000 breast cancer patients, from METABRIC testing set. The selected gene signature provides a powerful tool for the identification of patients at high-risk of death. We also describe how to extend the proposed approaches to integrate an additional omic layer, such as copy number aberrations, and we demonstrate that such strategies can further improve our prediction capabilities. We stress that although the retrieved signatures are specific for breast cancer survival, the proposed approaches can be used for different types of cancers.

More precisely, we propose new multistage computational-statistical strategies for survival analysis based on the following steps (see **Figure 1**). First, we reduce the high-dimensionality of data by using one of the following types of dimension reduction techniques: (i) a biomedical-driven screening (BMD-screening); (ii) a data-driven screening (DAD-screening); (iii) a combination of BMD-and-DAD-screening (BMD+DAD-screening). These screening approaches have different advantages and drawbacks. The BMD-screening is achieved by incorporating in the model the biomedical knowledge available in literature about breast cancer and, obviously, it can be performed only when there is enough evidence available. Nowadays such information can be often retrieved for previous studies and public databases, although it is still far from being complete. On the contrary, the DAD-screening relies only on the observed data. Therefore, it is suggested when there is limited biomedical information available. To fill the gap between the two approaches, the BMD+DAD-screening is introduced to take advantage of the available biomedical knowledge and also to allow finding novel elements of investigation that can emerge from data. Hence, the BMD+DAD screening can be used when there is partial biological information available and novel information is expected to be present in the data under analysis. Such situation represents the most common case. Second, we used penalized Cox regression methods (such as AdaLnet and ADMMnet) to incorporate gene regulatory relationships and to select a subset of potential biomarkers. In carrying out this step we show that when the BMD+DAD-screening is used we detect novel disease risk genes that the simple BMD-screening ignores. Third, we validate the proposed procedure and we evaluate the predictive power of the selected gene signatures. Finally, we perform a

pathway analysis based on the screened genes using Human Experimental/Functional Mapper (Huttenhower et al., 2009), KEGG pathways and COSMIC to make sense of the potential biomarkers for breast cancer survival. Moreover, to illustrate the advantages of multi omic integration, we first compare the performance of our approaches using only gene expression data, and then we integrate both gene expression and copy number aberrations. Our analysis shows that the integration provides a more comprehensive picture of breast cancer and improves the results.

Before concluding, we observe that in our previous work (Iuliano et al., 2016), we proposed a similar strategy based on the use of biological network into Cox regression penalized methods. That approach was similar in the spirit to the BMD-screening, discussed here. While confirming previous results using an independent dataset, the novelty of this article consists in the use of DAD and BMD+DAD screenings that allow us to extend and/or improve the performance of the proposed strategy, in the identification of novel potential biomarkers, and in the possibility of integrating multiple omic data types in a comprehensive analysis.

2. METHODS

In this section, (i) we introduce the Cox proportional hazards model (Cox, 1972); (ii) we present the three different types of screening techniques used to reduce the feature space to a subset of significant variables; (iii) we discuss network-regularized methods for selecting gene signatures; (iv) we describe the proposed algorithm; (v) we illustrate the extension of the algorithm for the integration of two omic data layers; (vi) we show how to make sense of the retrieved gene signature by using pathway analysis, and finally (vii) we discuss details about the implementation of our algorithm.

2.1. Cox Proportional Hazards Model

Let n be the number of subjects (patients), T_i and C_i for $i = 1, \ldots, n$ the survival time and the censoring time, respectively. Moreover, we denote the observed survival time as $t_i = min\{T_i, C_i\}$, the censoring indicator as $\delta_i = I(T_i \leq C_i)$ [where $I(\cdot)$ represents the indicator function], the regressor vector of p-variables for the ith subject (i.e. multi-omics observed profiles of the ith patient over p genes) as $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$. We also assume that the survival time T_i and the censoring time C_i are conditionally independent given the regressors $\mathbf{x}_i$ and the censoring mechanism is noninformative. Hence, the observed data are represented by the triplets $\{(t_i, \delta_i, \mathbf{x}_i), i = 1, ..., n\}$.

Under the assumption of Cox regression (Cox, 1972) the hazard function $h(t|\mathbf{x}_i)$ can be written as

$$h(t|\mathbf{x}_i) = h_0(t) \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)$$

where $h_0(t)$ represents the baseline hazard and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ the vector of regression coefficients. In the classical settings, the regression parameters are estimated by maximizing the Cox's log-partial likelihood

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

where $R(t_i)$ denotes the risk set at time t_i (i.e., the set of all patients who still survived prior to time t_i).

When the number of genes p is much larger than the patients n ($p \gg n$), such approach cannot be applied since the solution become not identifiable. To cope with this issue, improving prediction performance and the interpretation of the data, several penalization approaches have been proposed. Such techniques consist in adding a ℓ_1-penalty and/or ℓ_2 penalty term to the log-likelihood (1) in order to reduce the solution space imposing sparsity and small coefficients for the parameters (Tibshirani, 1996; Tibshirani et al., 1997; Zou and Hastie, 2005).

2.2. Variable Screenings for Cox's Proportional Hazards Model

The first step of our strategy is the variable screening of the data which aimed to reduce the number of variables for a large to a moderate scale. To this purpose, we assume that only a small number of these p variables is affecting the survival outcome. Therefore, we filter out variables that are considered not relevant for the disease under investigation. To this purpose, we consider three different types of variable screenings: biomedical screening (BMD-screening), data-driven screening (DAD-screening) and the fusion of biomedical and data-driven screening (BMD+DAD-screening). In the following sections, we define the set $\{x_j, j \in \mathcal{I}\}$ as the subset of the screened variables and $d = |\{x_j, j \in \mathcal{I}\}|$ its cardinality.

2.2.1. Biomedical-Driven Screening

In this type of screening to identify the subset $\{x_j, j \in \mathcal{I}\}$ we used only the biological information that has been accumulated in the literature on the cancer disease under investigation (Iuliano et al., 2016) and it is available in some external databases. In particular, as source of biological information (i.e., genes potentially associated to breast cancer) we used Human Experimental/Functional Mapper (HEFaIMp) (Huttenhower et al., 2009). Such web-resource describes the genes functional activity and gene-gene interactions in over 200 areas of human cellular biology with information from 30,000 genome-scale experiments and summarizes information from different biological informative sources such as prediction of protein function and functional modules, cross-talk among biological processes, and association of novel genes and pathways with known genetic disorders. HEFaIMp provides a p-value for each gene that indicates how significant is the relation between the gene and the disease of interest. Hence, we define with $\mathcal{I}_{BMD}$ the subset of genes selected by using HEFaIMp tool with p-value less or equal than 0.05 and with d_{BMD} its cardinality. We called this screening BMD-screening.

Note that, in the BMD-screening, we select the $\mathcal{I}_{BMD}$ set using standard p-value (with significance threshold equal to 0.05) without controlling for multiple tests, because we use a two-stage procedure composed by a screening step followed by a variable

selection method (i.e., the network approach described in 2.3). In this context, the identification of the variables associated to our pathology is performed in the variable selection step, and the screening is simply aimed to perform a pre-selection of the features. Such approach is typically done in the context of statistical screenings (see Fan and Lv, 2008). However, to further screen the variables of interest it could be also possible to control the multiplicity at the screening level, as described in Dmitrienko et al. (2009).

2.2.2. Data-Driven Screening

In this type of screening to identify $\mathcal{I}_{DAD}$ we reduce the feature space from a large scale dimension p to a relatively moderate scale $d < p$ by using only information from the data. This type of knowledge consist of the matrix $\mathbf{X}$ that contains single omic or multi omics patient profiles. Such approach differ from the BMD-screening where the information on which gene filter out and which retain in the model was obtained from an external database.

Let $\mathcal{M}_* = \left\{1 \leq i \leq p : \beta_i^* \neq 0\right\}$ be the true sparse Cox model. The maximum marginal likelihood estimator (MMLE) β_k^M, for $k = 1, \ldots, p$, is defined in Cox model as the maximizer of the log-partial likelihood with a single covariate

$$\beta_k^M = \arg\max_{\beta_k} \sum_{i=1}^{n} \delta_i \left\{ x_{ki}\beta_k - \log\left[\sum_{j \in R(t_i)} \exp(x_{kj}\beta_k) \right] \right\}. \tag{2}$$

The component-wise estimators can be computed very rapidly and implemented modularly, avoiding the numerical instability associated with ultrahigh dimensional estimation problems. The SIS procedure ranks the importance of features according to the magnitude of their marginal regression coefficients. Therefore, we select a set of variables

$$\mathcal{I}_{DAD} = \{1 \leq k \leq p : |\beta_k^M| \geq \delta_n\} \tag{3}$$

where δ_n is a threshold value chosen so that we pick the d_{DAD} top ranked covariates. The higher correlation, the higher the ranking position. As often suggested, one may choose $\lfloor n \log n \rfloor$ as threshold to select the most appropriate number of genes to retain in the model. More in general, the choice of the threshold may also be either data-driven or model-based. However, the aim of the screening procedure is to filter out as many noisy variables as possible, retaining all interesting ones in the model. After that the penalty in the network-based approach will select the few most relevant features.

For this reason, in our study we select different thresholds and we study their effect to optimize data prediction. It is easy to note that larger d_{DAD} means larger probability of including the true model $\mathcal{M}_*$ in the final model with indices in $\mathcal{I}_{DAD}$. We called this screening DAD-screening.

2.2.3. Biomedical-Driven and Data-Driven Screening

In this type of screening to identify $\mathcal{I}_{BMD+DAD}$ we merge the biological information known in literature and the data-driven knowledge to obtain new insights about cancer diseases by taking the union of the BMD and DAD sets of genes, i.e., $d_{BMD+DAD} = d_{BMD} \cup d_{DAD}$. Indeed, no cancer has been yet fully characterized in term of disrupted genes and/or metabolic processes involved in the disease. In particular, the BMD and DAD screenings take into account respectively available biological knowledge (i.e., genes highly correlated to breast cancer as described in the literature) and genes closely associated with the survival response (as emerging from the data). The BMD and DAD screening represent two faces of the same medal and naturally complement themselves. By using BMD+DAD screening, we aim to explore the best model that can sufficiently explain the data in the most parsimonious way in order 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. In fact, although breast cancer is the most common cancer types analyzed in literature, still remains a need for a more comprehensive and exhaustive study to find and investigate novel biomarkers. We called this screening BMD+DAD-screening.

2.3. Network Approaches After Screening

The second step of our strategy is the application of penalized methods using the subset of screened variables $\{x_j, j \in \mathcal{I}\}$ (where $\mathcal{I}$ depends on the type screening performed) as new feature space to further remove not significant variables from the model. The Cox penalized partial likelihood is

$$\ell(\boldsymbol{\beta}_{\mathcal{I}}) = \arg\min_{\beta_{\mathcal{I}}} \left(\sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_{\mathcal{I},i}^T \boldsymbol{\beta}_{\mathcal{I}} - \log\left[\sum_{j \in R(t_i)} \exp(\mathbf{x}_{\mathcal{I},j}^T \boldsymbol{\beta}_{\mathcal{I}}) \right] \right\} + P_\lambda(\boldsymbol{\beta}_{\mathcal{I}}) \right), \tag{4}$$

where $\boldsymbol{x}_{\mathcal{I},i}^T$ denotes a sub-vector of $\mathbf{x}_i$ with indices in $\mathcal{I}$, see Equation (3). $\boldsymbol{\beta}_{\mathcal{I}}$ are the screened regression coefficients. In particular, we add a penalty function $P_\lambda(\boldsymbol{\beta}_{\mathcal{I}})$ on the regression coefficients $\boldsymbol{\beta}_{\mathcal{I}}$. In the following section we introduce network-penalized approaches on the screened genes $\{x_j, j \in \mathcal{I}\}$ to incorporate an a-priori biological knowledge into the model and to predict survival outcomes.

2.3.1. Network-Regularized Cox Regression

The existing relationships among the covariates can be described in terms of a weighted and undirected graph (network) $G = (V, E, W)$ where the vertices $V = \left\{1, \ldots, d\right\}$ represents genes or covariates, an element (i, j) in the edge set $E \subset V \times V$ indicates a relationship between vertices i and j. $W = (w_{ij})$, $(i, j) \in E$ represent the weights (or strength of the relationship) associated with the corresponding edges. The relationships between genes can be obtained in terms of gene-gene interaction, KEGG pathway analysis or protein-protein interaction, or other functional information and it is normalized in $[0, 1]$ where 0 indicates an absence of relationship and 1 a strong relationship. More in general, the weight may indicate the probability that two genes are functionally connected. Such information is incorporated in the analysis using a penalty function $P_\lambda(\boldsymbol{\beta}_{\mathcal{I}})$ in Equation (4).

More formally, we introduce the following network penalty function

$$P_{\lambda,\alpha}(\boldsymbol{\beta}_{\mathcal{I}}) = \lambda\left[\alpha\,\|\boldsymbol{\beta}_{\mathcal{I}}\|_p + (1-\alpha)\,\Phi(\boldsymbol{\beta}_{\mathcal{I}})\right] \tag{5}$$

where $\lambda > 0$ (sparsity) and $\alpha \in (0,1]$ (network influence) are two regularization parameters (Zhang et al., 2013; Guo and Wan, 2014). The subset $\mathcal{I}$ includes the variables selected by using the previous screening approaches (BMD-screening, DAD-screening or BMD+DAD-screening). The penalty function is composed by two terms. The first part is a ℓ_p-norm with $p \in \{1,2\}$ which induces sparsity or thresholding; the second one $\Phi(\cdot)$ is a Laplacian matrix constraint which gives smoothness among two adjacent coefficients in the network. Generally, $\Phi(\cdot)$ for every pair of genes linked by an edge, which is proportional to the edge weight and the difference between their coefficients is a cost function. This hypothesis indicates that the two genes should be correlated. In other words, the regression coefficients should be similar, i.e., vary smoothly through the network (Zhang et al., 2013; Sun et al., 2014).

In our work, we use two of the most recent network-based Cox regression models. The details of each method are listed below.

The first method is based on a-priori network information is *Ada*ptive *L*aplacian *net* (or *AdaLnet*) (Sun et al., 2014). Denoting with $d_i = \sum_{i:(i,j)\in E} w_{ij}$ the degree of vertex i, *AdaLnet* defines the normalized Laplacian matrix $\mathbf{L} = (l_{ij})$ of the graph G (positive semi-definite) by

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

The network-constrained penalty in Equation (4) is given by

$$P_{\lambda,\alpha}(\boldsymbol{\beta}_{\mathcal{I}}) = \lambda\left[\alpha\,\|\boldsymbol{\beta}_{\mathcal{I}}\|_1 + (1-\alpha)\,\Phi(\boldsymbol{\beta}_{\mathcal{I}})\right], \tag{6}$$

where

$$\Phi(\boldsymbol{\beta}_{\mathcal{I}}) = \sum_{(i,j)\in E} w_{i,j}\left(\frac{sgn(\tilde{\beta}_{i,\mathcal{I}})\beta_{i,\mathcal{I}}}{\sqrt{d_i}} - \frac{sgn(\tilde{\beta}_{j,\mathcal{I}})\beta_{j,\mathcal{I}}}{\sqrt{d_j}}\right)^2.$$

The penalty in Equation (6) is the sum of an ℓ_1-penalty that brings sparsity and a quadratic Laplacian penalty that induces smoothness between adjacent vertices in the network. The vector $\tilde{\boldsymbol{\beta}}_{\mathcal{I}}$ is obtained from a preliminary regression analysis. The scaling of the coefficients $\boldsymbol{\beta}_{\mathcal{I}}$ 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 the penalty in Equation (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, here λ is the parameter that regularizes by the likelihood network constraint and $\alpha \in (0,1]$ is the parameter balancing the network constraint with respect to the sparsity.

The second network penalized method is based on the Alternating Direction Method of Multipliers (ADMM) algorithm used to solve a broad range of statistical optimization problems (Boyd et al., 2011). ADMM is an algorithm that solves convex optimization problems by breaking them into smaller pieces, each of which are then easier to handle. The algorithm solves problems in the form:

$$\text{minimize } f(\mathbf{x}) + g(\mathbf{z}) \quad \text{subject to } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c} \tag{7}$$

with $\mathbf{x} \in \mathbb{R}^{n\times 1}$, $\mathbf{z} \in \mathbb{R}^{m\times 1}$, $\mathbf{A} \in \mathbb{R}^{p\times n}$, $\mathbf{B} \in \mathbb{R}^{p\times m}$,and $\mathbf{c} \in \mathbb{R}^{p\times 1}$. The functions f and g are supposed convex. The optimal value of the problem Equation (7) will be denoted by

$$p^* = \inf\{f(\mathbf{x}) + g(\mathbf{z}) | \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}\}.$$

An alternative formulation is the following Lagrangian form

$$L_\rho(\mathbf{x},\mathbf{y},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^T(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{z}-\mathbf{c}) + (\rho/2)\|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{z}-\mathbf{c}\|_2^2 \tag{8}$$

ADMM consists of the iterations:

$$\begin{aligned} \mathbf{x}^{k+1} &:= \text{argmin}\, L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k), \\ \mathbf{z}^{k+1} &:= \text{argmin}\, L_\rho(\mathbf{x}^{k+1}, \mathbf{z}, \mathbf{y}^k), \\ \mathbf{y}^{k+1} &:= \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}). \end{aligned} \tag{9}$$

where $\rho > 0$. The algorithm consists of an x-minimization step, a z-minimization step, and a dual variable update (see Equation 9). Therefore, the method of multipliers for solving the problem in Equation (9) has the form

$$(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) := \text{argmin}_{\mathbf{x},\mathbf{z}}\, L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}^k)$$

$$\mathbf{y}^{k+1} := \mathbf{y}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c}).$$

The algorithm state in ADMM consists of $\mathbf{z}^k$ and $\mathbf{y}^k$, i.e. $(\mathbf{z}^{k+1}, \mathbf{y}^{k+1})$ is a function of $(\mathbf{z}^k, \mathbf{y}^k)$. In ADMM form, our problem can be written as

$$\text{minimize } f(\mathbf{x}) + g(\mathbf{z}) \quad \text{subject to } \mathbf{x} - \mathbf{z} = 0 \tag{10}$$

where $f(\mathbf{x}) = \ell(\boldsymbol{\beta}_{\mathcal{I}})$ (Equation 1) and $g(\mathbf{z}) = P_{\lambda,\alpha}(\boldsymbol{\beta}_{\mathcal{I}})$ (Equation 5 with $p = 1$) evaluated on $\boldsymbol{\beta}_{\mathcal{I}}$. The network penalty function is given by

$$P_{\lambda,\alpha}(\boldsymbol{\beta}_{\mathcal{I}}) = \lambda[\alpha\|\boldsymbol{\beta}_{\mathcal{I}}\|_1 + (1-\alpha)\phi(\boldsymbol{\beta}_{\mathcal{I}})],$$

where $\phi(\boldsymbol{\beta}_{\mathcal{I}}) = \boldsymbol{\beta}^T L \boldsymbol{\beta}$ with L is the Laplacian matrix.

2.3.2. Laplacian Matrix

The Laplacian matrix that describes the functional relationships among genes is constructed as done in our previous work (Iuliano et al., 2016) for genes covered by HEFalMp tool; the matrix is completed by adding zero weights for all the genes that are non-covered by HEFaIMp.

2.3.3. Tuning Parameters by k-Fold Cross-Validation

In principle, cross-validation can be used for estimating both α and λ. However the global procedure can be time-consuming and often provide only a limited improvement with respect to the optimization carried out using only one parameter. Therefore, we fixed $\alpha = 0.5$ and we estimate λ using cross-validation. To better understand the rationale of such choice, we note that $\alpha \in [0, 1]$ represents the influence of the network in the model. Small values of α will result in no network influence, large values α indicate strong influence. The choice $\alpha = 0.5$ assumes moderate influence of the network and represents a standard default parameter.

In order to estimate λ, the dataset is partitioned in $K = 5$ different folds, where four parts are used for finding model's coefficients $\hat{\boldsymbol{\beta}}^{(-k)}_{(\lambda,\alpha)}$ and one part is used for assessing the prediction on unseen data. This procedure is repeated 5 times, shuffling the folds. The estimate is obtained by maximizing the cross-validation log-partial likelihood (CVPL) defined as

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

where $\hat{\boldsymbol{\beta}}^{(-k)}(\cdot)$ is the estimate obtained from excluding the kth part of the data with a given pair of (λ, α), $\ell(\cdot)$ is the Cox log-partial likelihood on all the sample and $\ell^{(-k)}(\cdot)$ is the log-partial likelihood when the kth fold is left out (van Houwelingen et al., 2006). To assess the stability of the survival prediction we performed the five-fold cross-validation 10 times and we take as estimate the average value of λ.

2.3.4. Survival Analysis

The results of section 2.3.1 consist in a gene signature, i.e., $\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$, that can be used to predict patient survival. Survival analysis is performed using the Kaplan Meier curves after dividing the patients in two risk groups (high-and-low risk group) on the basis of the prognostic index computed with the gene signature. The p-value, used to test the null hypothesis that the survival curves are identical vs. the alternative that the two groups have different survival, is calculated by using the log-rank test.

2.4. General Algorithm for the Screening-Network Survival Prediction

In this section, we present the general procedure used for model's prediction.

Algorithm 1. *Screening-network survival prediction.*
Let define T the training set and D the validation set.

1. Apply screening techniques on T to reduce the dimension of the variable space from a large scale p to a moderate scale d, $d < p$. BMD- or DAD- or BMD+DAD-screening can be used for such purpose.
   (a) Define the subset $\{x_j, j \in \mathcal{I}\}$ as the subset of the screened variables.
2. Perform network-based Cox regression methods on $\{x_j, j \in \mathcal{I}\}$ in order to select the high-risk cancer genes. Either AdaLnet or ADMM can be used in this step.
   (a) Fix the regularization parameter $\alpha = 0.5$ to assess the network influence.
   (b) Repeat five-fold cross validation 10 times and take the mean of this estimate as the optimal tuning parameter values $(\hat{\lambda}_{\mathcal{I}}, \hat{\alpha}_{\mathcal{I}})$.
   (c) Use $\hat{\lambda}_{\mathcal{I}}$ and $\hat{\alpha}_{\mathcal{I}}$ to fit the corresponding penalized model and denote the parameter estimate by $\hat{\boldsymbol{\beta}}_{\mathcal{I}}$.
   (d) Select the BMD- or DAD- or BMD+DAD-genes with regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$.
3. Compute the prognostic index (PI) for each patient i in T, for $i = 1, \ldots, n$, as
$$PI_i^{\mathcal{I}} = \boldsymbol{x}_i^{\mathcal{I}}\hat{\boldsymbol{\beta}}_{\mathcal{I}}, \tag{11}$$
where $\boldsymbol{x}_i^{\mathcal{I}}$ is the vector of screened gene expression value (or adjusted expression) associated to the i-th patient.
   (a) $PI_i^{\mathcal{I}}$ is used to partition the patients in two subgroups, that correspond to the high-risk and low-risk prognosis groups, as follows:
      i. Compute the quantile q_γ of $PI_i^{\mathcal{I}}$, with $\gamma = 0.20, 0.25, 0.30 \ldots, 0.80$.
      ii. Each patient i in T is assigned to the high-risk (or low-risk) group if its prognostic index $PI_i^{\mathcal{I}}$ is above (or below) the q_γ-quantile.
      iii. The optimal cutoff $PI^{*,T}$ is selected adaptively on T. Here, the optimal cutoff is the γ-value that corresponds to the best separation in high-and-low risk group with respect to the log-rank test as defined in Iuliano et al. (2016).
4. Calculate the prognostic index PI_i^D by using $\hat{\boldsymbol{\beta}}_{\mathcal{I}}$ and $PI^{*,T}$.
   (a) Each patient i in D is assigned into the high/low-risk group if its prognostic index $PI_i^D = \boldsymbol{x}_i^{\mathcal{D}}\hat{\boldsymbol{\beta}}_{\mathcal{I}}$ is above (or below) the fixed threshold $PI^{*,T}$. The value $\boldsymbol{x}_i^{\mathcal{D}}$ is the vector of gene expression value associated to the i-th patient in D.
5. Perform the log-rank test to compare the survival curves between the patients in the high-risk and low-risk groups defined by the predicted risk scores PI_i^D.
   (a) The performance measure is the p-value of the test (the significance level was set at 5%, i.e., p-value< 0.05).

2.5. Multiomics Data Integration

In the above description the matrix $\boldsymbol{X}$ is usually a classical gene expression matrix. In order to integrate the information of an additional omic layer we use MANCIE (matrix analysis and normalization by concordant information enhancement) (Zang et al., 2016). MANCIE can be applied using two (column-matched) data matrices and adjusts one (main matrix) using the other (associated matrix) by identifying and reinforce the

concordant information in the two matrices and reducing the discordant information between them. The two data matrices must contain two omic-profiles on the same set of samples/patients. For example, one can measure the same omic profile using different experimental platforms or one can consider different omic types. The main matrix refers to the type of data that is considered more relevant whose values are returned "adjusted." In this study, MANCIE was used to adjust mRNA data matrix (main matrix) using copy number aberrations (CNAs), as the associated matrix. The resulting adjusted matrix was used in our algorithm in the case of two omics analysis.

2.6. Pathway Analysis

Using $\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$, we perform a pathway analysis based on KEGG database to make sense of the proposed signatures (http://www.kegg.jp/ or http://www.genome.jp/kegg/). Therefore, we associated to each gene the list of KEGG pathway in which it is annotated and the number of publications that relates it to breast cancer. We represents our results in terms of a network. In order to draw such networks we considered only the *not isolated* genes, where a gene g is said *not isolated* if $G \cap K \supsetneq \{g\}$ (G denoting a given set of genes and K a given KEGG pathway). Namely, g is *not isolated* if there is at least another gene $g' \in G$ belonging to the same pathways of g. In such cases g and g' will be connected by an edge that depend on the pathway K.

In this representation, each node in the network represents a gene and an edge between two nodes means that the corresponding genes belongs to the same KEGG pathway. In particular, we use different colors for different pathways and three colors to identify the type of screened gene: orange color for genes selected by HEFaIMp tool with p-value < 0.05, green color for genes selected by HEFaIMp tool with p-value > 0.05, purple color for genes that are not explored by HEFaIMp tool. Triangular-shaped nodes correspond to the genes that have already been identified in literature as breast-cancer associated genes. The latter step has been done using the database available in Cotterill (1999). The number of papers that associates such genes to breast cancer is also reported in the triangular nodes.

We also use the Catalog Of Somatic Mutations In Cancer (COSMIC, v84) (Forbes et al., 2010) for exploring the impact of somatic mutations in breast cancer. We downloaded COSMIC database from https://cancer.sanger.ac.uk/cosmic/download. We analyzed genes obtained by the DAD-screening and BMD+DAD-screening.

2.7. Implementation of the Algorithm

The statistical approach presented in **Figure 1** and described in Algorithm 1 has been implemented as a comprehensive `R` script that allows to execute all methods under the same `R` environment. The METABRIC gene expression profiles (Molecular Taxonomy of Breast Cancer International Consortium) were downloaded from the European Genome-phenome Archive (EGA). Access to datasets was approved by the specified Data Access Committee (DAC). The Illumina probes were annotated with the mappings from the Bioconductor package `illuminaHumanv4.db` (Dunning et al., 2015). Whereas, the METABRIC copy number aberrations CNAs data were downloaded by cBioPortal for Cancer Genomics (www.cbioportal.org).

For the BMD-screening, we select a subset of genes that are involved in breast cancer by using a functional map that summarize the most relevant interactions in the cancer area of interest (Huttenhower et al., 2009). This map is used to build the network-matrix and to identify the weight of the edges among genes.

We use AdaLnet method which is a pathwise algorithm for the Cox proportional hazards model, regularized by network penalty [combination of ℓ_1-penalty, $\|\boldsymbol{\beta}_{\mathcal{I}}\|_1$ and Laplacian matrix $\Phi(\boldsymbol{\beta}_{\mathcal{I}})$] (Simon et al., 2011; Sun et al., 2014). It is implemented in `Coxnet package` (version 0.2, 2015-03-21). ADMM is an algorithm implemented in the `ADMMnet package` (version 0.1, 2015-12-12). For each method we fix the regularization parameter $\alpha = 0.5$ and repeat five-fold cross validation 10 times. Then we take the mean of this estimate as the optimal tuning parameter values (see Algorithm 1). Then, `Survival` package in the `R` software is used to compare the Kaplan-Meier survival curves and to derive the significance p-value indicating the difference between two survival curves. For the integration of different omic profiles `MANCIE package` was used (version 1.4, 2016-03-02).

Pathways analysis has been carried out by using the KEGG database through an integrative `R` script. RCytoscape (www.bioconductor.org/packages/release/bioc/html/RCytoscape.html) has been used to draw the networks (Shannon et al., 2013). Note that all the scripts are available upon request from the first two authors.

3. RESULTS AND DISCUSSION

In this section, we present the results obtained using the proposed approaches using the METABRIC dataset. For such purpose, we divided the dataset in two parts, training set (T) and testing set (D) as described in section 3.1 We compared the three screening procedures (BMD, DAD and DAD+BMD) combined with the two network Cox regression methods (AdaLnet and ADMM) with respect to the subset of screened genes (i.e., $\{x_j, j \in \mathcal{I}\}$) and their cardinality d, the potential biomarkers identified (i.e., those with regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}} \neq 0$), and the survival prediction capabilities. The screened genes and the potential biomarkers were evaluated on the training set, the latter resulting in a gene signature able to subdivide patients in high and low risk groups. The prediction capabilities were evaluated by using Kaplan-Maier curves and log-rank tests on the testing set. After that the list of potential biomarkers underwent to a pathway analysis in order to provide a biological interpretation of the results and illustrate the relationship with already available biological information. In discussing the results, we first show those obtained by analyzing only mRNA expression data, then we show the improvement observed by integrating mRNAs and CNAs using MANCIE as described in section 2.5. Overall our results show that the BMD+DAD-screening is better than BMD or DAD in terms of predictive power (i.e., smaller p-value for the log-rank test on the testing set) for breast cancer survival patients and also

allows to identify as potential biomarkers few genes that the BMD screening ignores. Moreover, we also demonstrate that integrating two omic data types improves the predictions.

In the following, AdaLnet method is referred as Coxnet and ADMM is referred as ADMMnet, according to the R packages where they are implemented.

3.1. Data Availability

We used METABRIC data to evaluate the performance of our screening-network approach. This dataset contains clinical traits, mRNA expression data, CNAs profiles, and SNP genotypes derived from 1980 breast cancer samples (patients) (Curtis et al., 2012). In particular in our comparison, we use mRNA expression data downloaded from The European Genome phenome Archive (EGA) with number EGAS00000000083 and the copy number aberrations (CNAs) available on cBioPortal for Cancer Genomics (http://www.cbioportal.org/). The mRNAs data consist in a matrix containing 48,803 Illumina expression probes measured on the Illumina HT-12 v3 platform. The CNAs matrix is coded using value -2 to indicate homozygous deletion; value -1 to represent the hemizygous deletion; value 0 meaning neutral/no change; value 1 showing the gain; value 2 for high level of amplification. Both the matrices are normalized as discussed in Curtis et al. (2012). By using these data, we conducted two types of analysis based on (i) mRNA expression data and (ii) integration mRNA and CNAs.

As a first step, we divided the patients in two subsets: a training set T (997 samples) and testing set D (995 samples). When performing the analysis using only the mRNA expression data a total of 19,151 genes was retrieved from 48,803 Illumina expression probes by using a bioconductor annotation data package (Dunning et al., 2015). When performing the analysis integrating mRNA and CNAs information a total of 18,006 genes (containing both mRNA and CNAs information) was considered from 26,298 copy number features summarized at the gene level. A summary of METABRIC dataset is shown in **Table 1**.

Finally, the overall survival (OS) data related to the 1980 patients (long-term follow-up data) were downloaded from cBioPortal for Cancer Genomics (Q_1 = 60.78 months, Median = Q_2= 116.10 months, Q_3 = 184.90 months). In particular, the OS-status indicator was divided in *died of disease* (deceased=1), *living* (censored=0) and *died of other causes* (censored=0), respectively (Gao et al., 2013).

TABLE 1 | METABRIC dataset summary: mRNA expression dataset and the integration of mRNA data CNAs profiles (mRNA+CNAs).

Omics data	Training set (T)		Testing set (D)	
	Sample	# Genes	Sample	# Genes
mRNA	997	19,151	995	19,151
mRNA+CNAs	997	18,006	995	18,006

For each case, the samples are divided into two subsets: training set T e testing set D, respectively.

3.2. Screening-Network Analysis

First, we describes the results obtained using the BMD-screening. In order to select $\{x_j, j \in \mathcal{I}_{BMD}\}$ we used HEFalMp tool (http://hefalmp.princeton.edu/hefalmp) and we selected only those genes that in the HEFalMp tool have p-value < 0.05 for breast cancer association. In particular, the BMD-screening selected a total of $d_{BMD} = 528$ genes when using mRNA expression data and $d_{BMD} = 526$ genes when integrating mRNA and CNAs data (see, Table S1 for the screened gene lists). Such subsets of genes reflect the bio-medical knowledge about breast cancer markers available from previous studies. HEFalMp was also used to build the gene network to be used in the network-penalized Cox regression method. Then, the network-based Cox regression methods applied on the training dataset, T, allowed us to select high-risk genes or potential biomarkers (i.e., those with regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD}} \neq 0$). We denoted this gene signature as BMD-genes (see **Table 2** and Table S2). BMD-genes were used to compute the prognostic index of each patient and to classify them in low and high risk groups. An optimal cut-off for the prognostic index was estimated for such purpose. The significance of the BMD-gene lists was evaluated on the testing dataset, D, in terms of p-values of the log-rank test were novel patients were divided in low and high risk groups according to their prognostic index. **Figure 2** shows the Kaplan-Meier survival curves on the testing set D for each combination between the BMD-screening and the network-penalized Cox regression methods, by using only mRNA expression data and the integration between mRNA and CNA profiles, respectively. **Figures 2A,B** refer to Coxnet and **Figures 2C,D** to ADMMnet. **Table 2** shows additional results of our procedure in terms of identified markers in the training set T and log-rank test p-value obtained from the testing set D. Overall such results confirm those obtained in Iuliano et al. (2016) on an independent datasets. Moreover, they also show that the integration (mRNA+CNA data) of two omic types provides a better prediction of patient survival (i.e., better separation in terms of p-value) than the use of a single omic layer (mRNA expression data), thus extending the results of previous work.

To better understand the BMD-genes signature obtained using mRNA and/or mRNA+CNAs data, we show the heatmap of the gene expression. In particular, we ordered the patients with respect to the prognostic index PI and divide them in two risk classes (i.e., low-risk and high risk) using the optimal cut-off PI^*,

TABLE 2 | Number of BMD-genes selected by using the combination of BMD-screening and network-penalized Cox methods (Coxnet and ADMMnet) with regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD}} \neq 0$ on the training set T.

Omics data	Methods	# BMD-genes	p-value	α	λ
mRNA	Coxnet	38	1.6e-05	0.5	0.07934
	ADMMnet	43	8.12e-06	0.5	0.07695
mRNA+CNAs	Coxnet	24	1.09e-07	0.5	0.09338
	ADMMnet	19	3.3e-08	0.5	0.10170

The tuning parameters ($\lambda_{\mathcal{I}_{BMD}}, \alpha_{\mathcal{I}_{BMD}}$) and the relative p-values obtained from the testing set D are also shown.

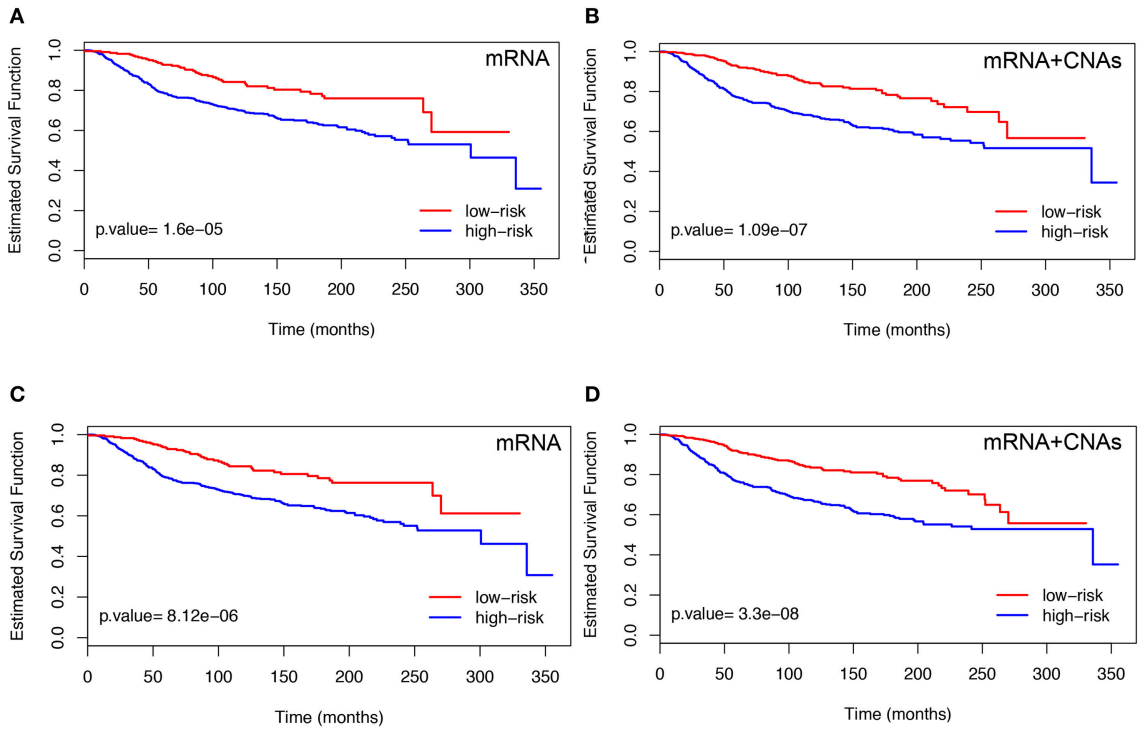

FIGURE 2 | Kaplan-Meier plots obtained using mRNA data (left) and the integration of mRNA and CNAs profiles (right). The results refer to the testing set *D*. For each case, patients were divided into two groups according to the prognostic index by using BMD-genes with Coxnet **(A,B)** and ADMMnet **(C,D)**. We use the color blue to indicate the high-risk group and the color red to show the low-risk group. The *p*-value is also calculated applying the log-rank test on testing set. The high-risk group is better separated from the low-risk group by using the integration of mRNA expression data and CNAs profiles (right), compared with using the single omics data (left). The X-axis represents time and the Y-axis represents survival rate.

as described in section 2.4. Figure S1 shows the Z-score matrix of the BMD-genes expression in the training (T) and testing (D) sets, respectively and Figure S2 shows similar heatmaps for the Z-score matrix of the adjusted BMD-genes expression. In each figure, the first row refers to the BMD-genes signature obtained using Coxnet, the second row using ADMMnet.

By inspecting the heatmaps in Figure S1, we identified two groups of genes (e.g., *PPDZK1, LRP2, PCM1, TMEM26, BCL2, AFF3*) and (e.g., *FUT3, FGFR4, CDC7, RRM2, SPC25, PKMYT1, UBE2C, TROAP*). The first group contains genes such that the lower is their expression the worse is the patient prognosis, the other group contains genes such that the higher is their expression the worse is the patient prognosis. There are however other genes for which the separation of the z-scores in the two risk groups is less evident, as already noticed also in Ahmad and Fröhlich (2017). Figure S2 shows similar behavior and group of genes, reducing the noise in the heatmaps. In this case we identified the same group of genes and few others of interest. Among the latter, for *AURKA* the higher is the expression the worse is the prognosis, as also shown in Jiang et al. (2010).

Second, we show the results obtained using the DAD-screening. In this case, to select $\{x_j, j \in \mathcal{I}_{DAD}\}$ we used the DAD screening to reduce the dimensionality of the full dataset from p to $d_{DAD} < p$, for different thresholds $d_{DAD} = 100, 200, \ldots, 2,000$. Then, as before, we further reduced the model size down to $d' < d_{DAD}$ by fitting a network-based methods for each fixed threshold d_{DAD}. We called DAD-genes the high-risk gene signature (i.e., those genes with regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{DAD}} \neq 0$). Different choices of the threshold $d_{DAD} = 100, 200, \ldots, 2,000$ lead us to slightly different, but usually overlapping, DAD-gene lists. As before, the significance of the DAD-gene lists were assessed on the testing dataset, D. From our analysis we observed that the log-rank test p-values were able to separate the high and low risk group of patients with a significance lower than 0.01 only for some range of thresholds. As expected, log-rank p-value associated to the DAD-genes are not as strong as the corresponding p-values associated to the BMD-genes, suggesting that DAD-screening is not competitive in terms of prediction power with respect to the BMD-screening. Therefore, the information available from the literature should not be neglected and DAD-screening should be used to find potential candidate biomarkers and predict survival only when no other (or very limited) information is available. Anyway, our results also show that the performance of DAD-screening improves when two integrated omic types (mRNA+CNAs) are used instead of the simple gene expression (mRNA) profiles.

Finally, we discuss the results obtained using the BMD+DAD-screening. In this case, to select $\{x_j, j \in \mathcal{I}_{BMD+DAD}\}$ we merge the two above mentioned-screenings $\{x_j, j \in \mathcal{I}_{BMD} \cup \mathcal{I}_{DAD}\}$ using different thresholds $d_{DAD} = 100, 200, \ldots, 2,000$ when adding

the DAD contribution. Such subsets of genes reflect the biomedical knowledge available from previous studies (BMD part) and also incorporate additional information contained in the data under analysis (DAD part). Analogously to the previous cases, we fitted a network-based Cox regression model in order to further reduce the feature space from $d_{BMD+DAD}$ to d' and to select the high-risk genes or potential biomarkers (i.e., genes with regression coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{I}_{BMD+DAD}} \neq 0$) . We called this signature BMD+DAD-genes. As before, the significance of the BMD+DAD-gene lists was evaluated on the testing dataset, D, in terms of p-values of the log-rank test for each value of the threshold.

Moreover, in order to understand the BMD and the DAD contribution to the BMD+DAD-genes we subdivided the BMD+DAD-genes in:

a. genes-HEFaIMp-high: BMD+DAD-genes that match the genes selected by HEFaIMp tool with p-value< 0.05;
b. genes-HEFaIMp-low: BMD+DAD-genes that match the genes selected by HEFaIMp tool with p-value> 0.05;
c. genes-no-HEFaIMp: BMD+DAD-genes that are not covered by HEFaIMp tool.

Genes in group (a) are those included in the BMD-screening; genes in group (b) are presented in HEFaIMp but their evidence was not sufficiently strong to let them be included in the BMD-screening. However, our analysis reinforce the evidence that they could be related to breast cancer. By contrast, genes identified in group (c) might be important for the process of novel biomarker discovery since they represent potential biomarkers not previously identified as associated to breast cancer.

Tables S3, S4 show the results obtained from the combination of BMD+DAD-screening and network-penalized methods (Coxnet and ADMMnet) for different thresholds $d_{DAD} = 100, 200, \ldots, 2,000$. From these results, we observed that the log-rank test p-value associated to the BMD+DAD-genes on the testing dataset is better (i.e., smaller) than the corresponding p-value obtained using the BMD-genes and DAD-genes in both cases investigated (mRNA and mRNA+CNAs data). Therefore, the BMD+DAD-screening outperforms the other two screenings allowing: (i) better separation between high-and-low-risk groups and (ii) identification of novel potential biomarkers. Moreover, our results also confirm that our prediction capability further improves when two omic layers (mRNA + CNAs) are used instead of a single omic layer (mRNA). See also Figure S3 for the combination of BMD+DAD-screening and Coxnet and Figure S4 for the fusion of BMD+DAD-screening and ADMMnet.

Then, Tables S5, S6 show the list of BMD+DAD-genes selected from each screening-network approach by using mRNA expression data and the integration of mRNA and CNAs data, respectively. Tables S5, S6 also show the number of times each gene in the signature was selected when changing the threshold and the network methods. We observed that the BMD+DAD-genes create a consensus gene-set signature that is quite robust with respect to the choice of the threshold and can be potentially highly associated with breast cancer prognosis. In particular, *AFF3*, *ARVCF*, *AURKA*, *BCL2*, *C17orf78*, *EXPH5*, *FEZF2*, *FGFR4*, *FUT3*, *LRP2*, *PDZK1*, *PKMYT1*, *REL*, *SPC25*, *TMEM26*, *TROAP*, *UBE2C* were identified by using both mRNA expression data and mRNA+CNAs data. For these genes the frequency of the occurrence is equal to 20 corresponding to the number of threshold used in our analysis. Finally, to further evaluate the robustness of gene signatures we used Venn diagrams (see **Figure 3**). From this figure we observed that the overlaps between screening and network methods is quite good, although there are specificities that explain the better performance of one combination with respect to another. Moreover, **Figure 3** also show that the BMD+DAD-screening selects novel potential disease risk genes that the simple BMD-screening ignores.

A more comprehensive analysis of these candidate genes is described out in the following section.

3.3. Pathway Exploration

In order to better understand and interpret the inferred gene signatures, in this section we report the results of the KEGG pathways analysis performed on the not-isolated genes in the signature (as described in section 2.6). To this purpose we considered the BMD+DAD gene lists identified using Coxnet and ADMMnet models with both mRNA data and the integration of mRNA+CNAs to build the final pathway networks reported in **Figures 4**, **5**. We used such networks to easily visualize the gene-gene interactions and the KEGG pathways involved in such interactions. Each node corresponds to a gene and the edges represent the KEGG pathways shared by the linked genes. Different colors for nodes have been used to indicate genes-HEFaIMp-high (orange), genes-HEFaIMp-low (green) or genes-no-HEFaIMp (purple) as defined in section 3.2. Therefore orange nodes represent the BMD contribution to the signature and green and purple nodes the DAD contribution, not yet captured in the BMD list. Note that some of the genes colored in orange might be also be retrieved from the data under analysis (as DAD-genes), however in this context we want to underline and make sense of the novel information not yet considered.

Figure 4 shows the gene-networks built on the genes identified by Coxnet and ADMMnet respectively using mRNA data. From the color of the nodes, we can infer that most of (but not all) the genes come from the BMD contribution (i.e., orange nodes). Hence, confirming that the BMD+DAD screening allows us to identify few genes that the BMD screening ignores. Moreover, our analysis allows us to further investigate the KEGG pathways the involved genes belong to. In particular, a gene shown in both networks is *BCL2*, which accordingly to Cotterill (1999) has already been mentioned in 122 publications showing its importance in breast cancer. *BCL2* functions to prevent apoptosis and it is a tumor-related gene that has the potential to further improve individualization of patient management, by predicting response to chemotherapy, hormonal therapy and radiotherapy (Joensuu et al., 1994; Hamilton and Piccart, 2000). In addition, as showed in both the networks in **Figure 4**, *BCL2* is involved in the *KEGG focal adhesion* pathway together with *MYL12B* and *ERBB2*. Extensive studies relate the *KEGG focal adhesion* pathway to breast cancer since it plays critical roles in integrin-mediated signal transduction and also participates in signaling by other cell surface receptors. *KEGG focal adhesion* pathway is also involved in angiogenesis during embryonic

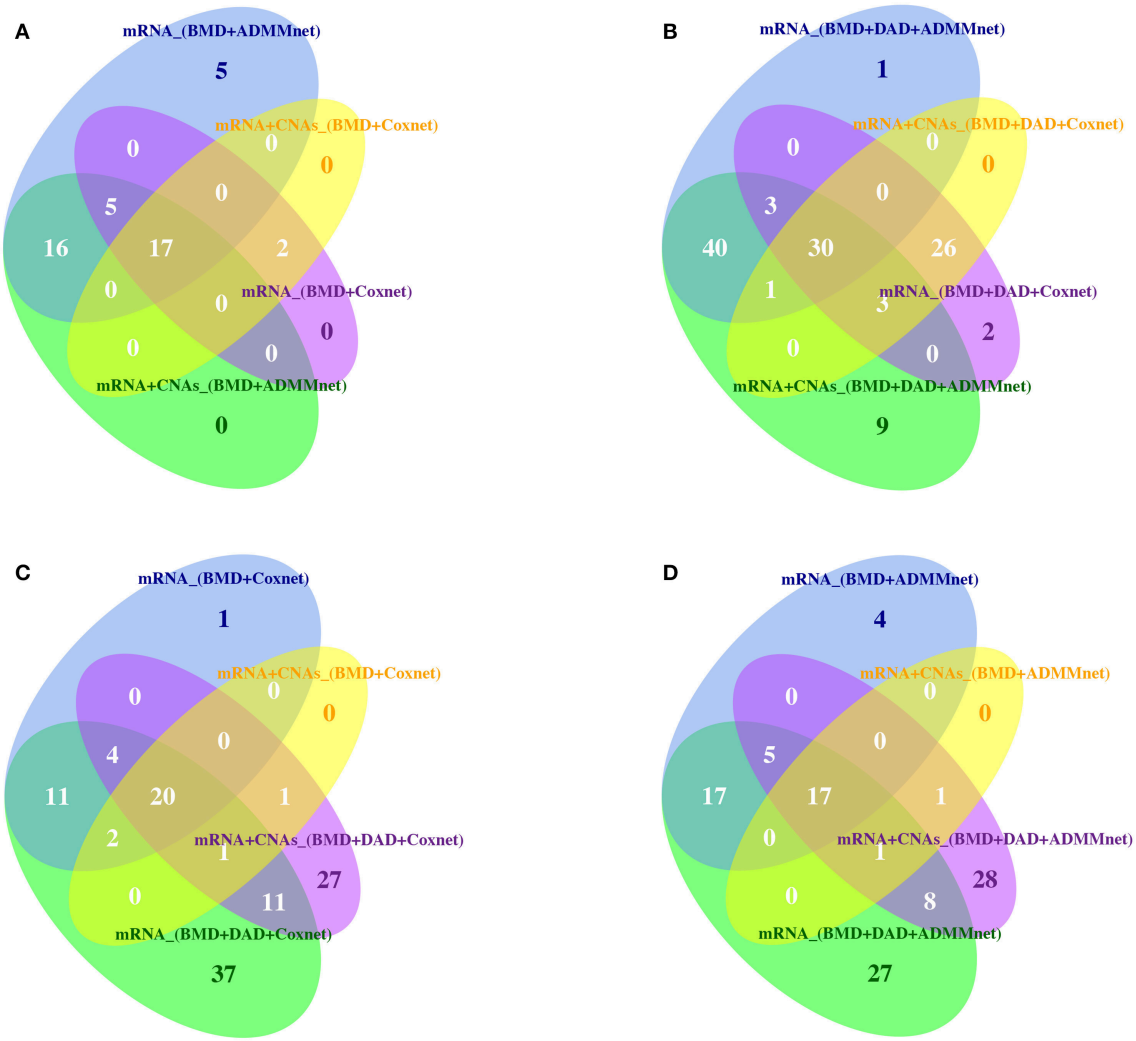

FIGURE 3 | Venn diagrams are used to illustrate **(A)** the intersection of BMD-genes by using Coxnet and ADMMnet with mRNA expression data and the integration of mRNA data and CNAs profiles; **(B)** the intersection of BMD+DAD-genes by using Coxnet and ADMMnet with mRNA expression data and the integration of mRNA data and CNAs profiles; **(C)** the intersection of BMD-genes and BMD+DAD-genes by using Coxnet with mRNA expression data and the integration of mRNA data and CNAs profiles; **(D)** the intersection of BMD-genes and BMD+DAD-genes by using ADMMnet with mRNA expression data and the integration of mRNA data and CNAs profiles.

development and cancer progression (Parsons, 2003; Cohen and Guan, 2005). In Zhao and Guan (2011), the authors also show the role of this pathway in cells migration and metastatic breast cancer. From the color of the three genes involved in this pathway (in both the networks in **Figure 4**), it results that even if *MYL12B* is not in the genes-HEFaIMp-high list (the node color is purple), it can play an important role in breast cancer. Indeed, *MYL12B* is involved in the regulation of cell morphology and recent studies have shown the link between such gene and cancer progression (Gurda et al., 2015). Another relevant gene reported in both networks is the fibroblast growth factor receptor-4 (*FGFR4*), which has been widely investigated as one of the major causes of disease progression in estrogen- and progesterone-receptor-positive tumors and in tumors with high lymph-node involvement (Jaakkola et al., 1993), confirming its relationship with breast cancer. Other cancer biomarkers have been reported in both networks with exactly the same pathway edges, which underline their important role in the disease and the accuracy of our algorithm. For example, *CDC7* and *PKMYT1* belong to the *KEGG cell cycle* pathways which is one of the most commonly disrupted pathways in cancer (Chang et al., 2003; Kastan and Bartek, 2004). Similarly *RRM2* and *ADCY2* belong to the *KEGG purine metabolism* pathway whose disruption is often linked with transformation and progression of cancer (Weber, 1983; Pedley and Benkovic, 2017). Overall, our results show that from the pathway analysis of the gene signatures using mRNA data, it is possible to investigate not only the genes involved in the progression of the

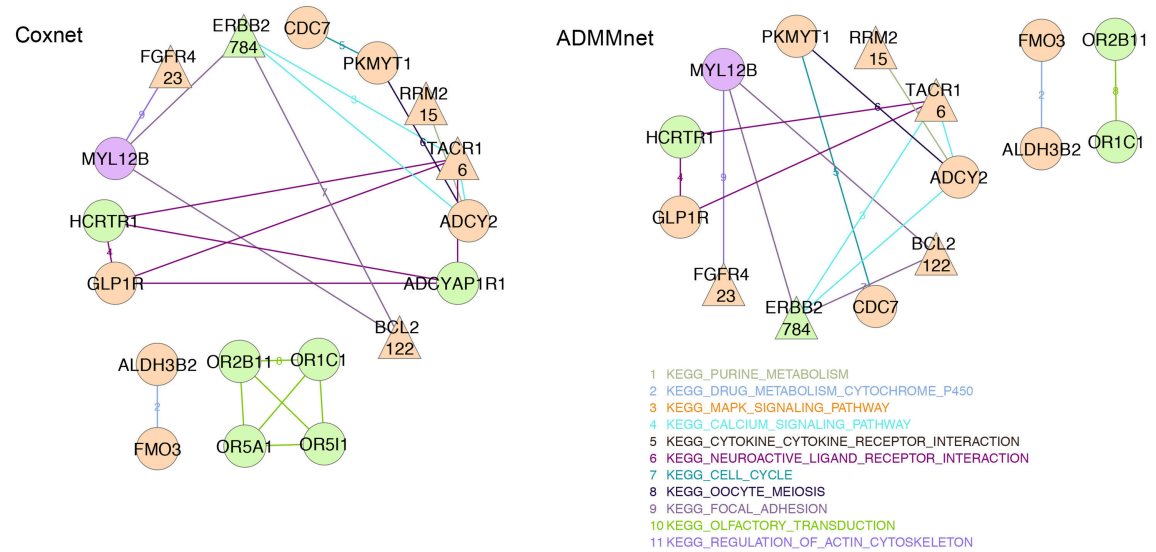

FIGURE 4 | mRNA expression data: BMD+DAD-screening network using Coxnet (left) and ADMMnet (right) methods with mRNA expression data and the integration of mRNA data and CNAs profiles. Non isolated genes are represented as nodes in the network, then a link a drawn between two (adjacent) genes when the two genes belong to the same KEGG pathway. We use different colors for KEGG pathways and three colors to identify the type gene: orange color for genes-HEFalMp-high, green color for genes-HEFalMp-low, purple color for genes-no-HEFalMp. Triangular-shaped nodes indicate the genes identified in literature as breast-cancer associated genes. The number of papers is also reported in the triangular nodes.

disease but also the relative pathways which may include novel biomarkers.

Figure 5 shows the networks corresponding to the genes identified by Coxnet and ADMMnet respectively using the integration of mRNA data and CNAs values. The majority of the nodes identified in those networks are green (i.e., with p-value > 0.05) which means that by using integrated data both Cox-methods select an higher number of DAD genes than before. It is worthy to note that the two networks are almost identical except for Coxnet that selects one more gene (*OR6YI*) from the *KEGG olfactory transduction* pathway (see **Figure 5**, Coxnet). This pathway has a functional role in the development and/or progression of melanoma and it may even contribute to tumorigenesis (Ranzani et al., 2017). Both networks report *OR5I1*, *OR10AD10*, and *OR8S1* as part of the *KEGG olfactory transduction* pathway and they are olfactory receptors that have been linked with the promotion of cancer cell invasiveness and metastasis emergence (Sanz et al., 2014). The *KEGG neuroactive ligand receptor interaction pathway* has been identified in all the four networks (**Figure 5**). However, a new gene is reported in both the data integration networks that was not reported in the mRNA network, i.e., *GLRA1*. Such gene has been mentioned in several cancer studies as involved in cancer development (Murakami and Hirano, 2008; Kreisler et al., 2010). Two new pathways have been identified by using the integrated data: *KEGG cytokine cytokine receptor interaction* pathway and *KEGG mapk signaling* pathway. Both pathways are essential for cancer-immune evasion in melanoma cells since they regulate a variety of cellular activities including proliferation, differentiation, survival, and death (Lin and Karin, 2007; Kim and Choi, 2010).

In conclusion, we can confirm that by using either mRNA data or the integration of mRNA data and CNAs values, our algorithm is able to identify genes already known to be associated with breast cancer as well as new potential candidate markers and disrupted pathways. As a consequence, either methods can be used for the analysis of cancer pathways depending on the availability of the biological information about the disease under investigation.

3.4. Further Analysis of Potential Breast Cancer Biomarkers

In order to further exploit the relevance of the potential novel biomarkers we have identified with our analysis, we present a gene enrichment integrating somatic mutation using the genes-HEFaIMp-low and no-genes-HEFaIMp lists. Our aim is to better understand the biological relevance and make sense of the genes that were ignored when using the BMD-screening, but were found significant using the BMD+DAD-screening. More precisely, we match the two sublists of BMD+DAD-genes with the Catalog Of Somatic Mutations In Cancer COSMIC (Forbes et al., 2010).

Exploring this additional source of information, some high-risk mutated genes (variant type-missense mutations) in breast cancer are retrieved. Among genes-HEFaIMp-low, an interesting protein coding gene is *EXPH5* (Exophilin 5). This protein has been identified as an important prognostic gene for breast cancer. In particular, this gene is connected to the missense variant (GAG → GTG) which is associated with differential methylation, gene expression, and survival of TCGA breast cancer patients (Shilpi et al., 2017). Another important protein coding gene is *FGFR4* (Fibroblast Growth Factor Receptor 4) which is an

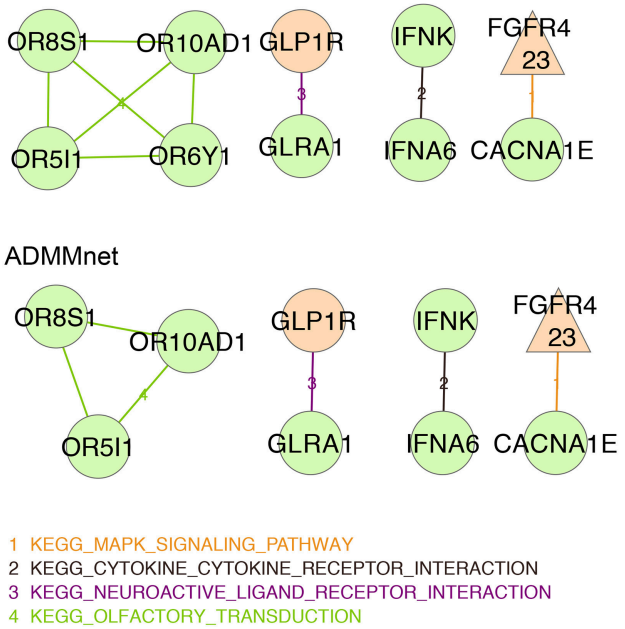

FIGURE 5 | Integration of mRNA data and CNAs profiles: BMD+DAD-screening network using Coxnet (top) and ADMMnet (bottom) methods with mRNA expression data and the integration of mRNA data and CNAs profiles. Non isolated genes are represented as nodes in the network, then a link a drawn between two (adjacent) genes when the two genes belong to the same KEGG pathway. We use different colors for KEGG pathways and three colors to identify the type gene: orange color for genes-HEFalMp-high, green color for genes-HEFalMp-low, purple color for genes-no-HEFalMp. Triangular-shaped nodes indicate the genes identified in literature as breast-cancer associated genes. The number of papers is also reported in the triangular nodes.

essential kinase critical for the proliferation and survival of basal-like breast cancer cells. In particular, this gene mediates cancer cell survival via the activation of PI3K/AKT signaling. Moreover, FGFR4 and FGF19 autocrine signaling may serve as a novel potential therapeutic target for the treatment of refractory basal-like breast cancers (Tiong et al., 2016). *GLP1R* (Glucagon Like Peptide 1 Receptor) is a further protein coding gene expressed in human breast cancer tissue. In particular, the activation of *GLP1R* attenuates breast cancer cells proliferation by inhibiting NF-κB activation and target gene expression (Hirata et al., 2013). Moreover, the protein coding gene *MSX2* (Msh Homeobox 2) is also implicated in breast cancer. It is an important regulator of melanoma cell invasion and survival. Its cytoplasmic expression was identified as prognostic biomarker in malignant melanoma patients (Gremel et al., 2011). Finally, the protein coding gene *TMEM26* (Transmembrane Protein 26) is another important gene expressed in ERα-positive and -negative breast cancer cell lines. In particular, patients who received aromatase inhibitor treatment tend to have a higher risk of recurrence when the expression of *TMEM26* is low. Moreover, *TMEM26* negatively regulates the expression of integrin β1, which is an important factor involved in endocrine resistance (Nass et al., 2016). Among no-genes-HEFaIMp, an important protein coding gene is *ACTL9* (Actin Like 9). An important paralog of this gene is ACTL7A which is implicated in diverse cellular processes, including vesicular transport, spindle orientation, nuclear migration, and chromatin remodeling. In particular, this gene is involved in a risk locus for breast cancer at 9q31.2 (chromosomal position) that provide evidence of an association between variants mapping to 6q25.1 (chromosomal position) and breast cancer risk in subjects of European ancestry (Fletcher et al., 2011). Another important protein coding gene is MYL12B (Myosin Light Chain 12B). Myosin regulatory subunit plays an important role in regulation of both smooth muscle and nonmuscle cell contractile activity via its phosphorylation and it is implicated in cytokinesis, receptor capping, and cell locomotion. In particular, it is predominantly expressed in Triple-Negative Breast Cancer (Ziegler et al., 2014). Among its related pathways there are the *Semaphorin interactions* and *Focal Adhesion* pathways. An important paralog of this gene is MYL12A. The protein coding gene SLC22A25 is also detected by COSMIC. This gene has been identified as hub gene into the mechanisms of gene regulation during breast cancer (Emmert-Streib et al., 2014). An important paralog of this gene is SLC22A9.

4. CONCLUSIONS

In this work, we combine variable screening procedures and network-penalized Cox models for high-dimensional survival data aimed to reduce the size of initial dataset to a moderate size and to determine pathway structures and potential biomarkers involved in cancer progression. By using these approaches, it is possible to obtain a deeper insight of the gene-regulatory networks and investigate the gene signatures related to the breast cancer survival time in order to understand how patient molecular features can influence survival in cancer. Breast cancer is used as illustrative example, however the proposed methods can be used for different types of cancers.

We illustrate the capabilities of our approaches to predict patient survival using METABRIC dataset. First we used one out of three different screenings methods: biomedical driven screening, data-driven screening and a combination of the two. Then, using the biological network, as prior information network, we performed network-based Cox model to identify specific signatures of genes and the corresponding pathways associated to breast cancer prognosis. Finally we used Kaplan-Meier curve and log-rank test to validate the goodness of the prediction. Hence, while the screening methods recruit the features with the best marginal utility to reduce the dimensionality of the data, the network incorporates the pathway information used as a prior knowledge network into the survival analysis. Overall, we can conclude that (i) the BMD-screening confirms previous results on independent dataset (Iuliano et al., 2016); (ii) the DAD-screening shows good performance in absence of any previous information but it is sub optimal with respect to the BMD-screening; (iii) the BMD+DAD-screening allows to discover novel potential biomarkers for breast cancer

that are disregarded by the BMD-screening and improve the BMD-screening in terms of prediction capabilities. Moreover, we also illustrate how to extend the proposed methodologies, initially sought for gene expression data, to the case when two omic data types are available on the same set of patients. In particular, we compared the results obtained by our procedures using only mRNA expression values with those obtained by integrating mRNAs and CNAs. From our results, we can conclude that the use of two omic layers always outperforms the results obtained with a single omic.

Finally, we investigated the potential relevance of the BMD+DAD-genes we have detected. Our results show that they are often connected known cancer genes and are significantly enriched in biological processes and pathways that are involved in breast cancer, or annotated in cancer mutations databases such as COSMIC. Although this computational analysis does not guarantee that such genes can be considered biomarkers, they make sense of biological processes involved in breast cancer progression and provide a strong suggestion toward the need of future studies for their biological validation.

It is clear that the proposed procedures can be applied to different cancer types to obtain a more accurate investigation of the development and progression of the disease. In fact, from one hand breast cancer represents one of the types of cancer for which there is a wide knowledge accumulated in the literature. Nevertheless, the BMD+DAD-screening shows that there is still space for improvements and for novel discoveries. On the other hand, the information available for some types of cancers might not be so accurate. Therefore, methods such as DAD-screening might be useful to provide a good level of analysis.

The results obtained in this work open interesting scenarios for future developments. First, we have shown that the use of two omic layers improves prediction capabilities, therefore the integration of data from multiple omics (e.g., structural variations, methylation or other epigenetic markers and/or metabolomics) into the screening procedure could also provide a more accurate investigation and prevent the limitations of current methods. The possibility of combine together different types of omics or other co-data is expected to further improve the results. Second, in order to support clinicians with a more concrete and biomedical perspective, the proposed procedures should be further extended in order to include also clinical and therapeutical information for each patients. Such information will allow to better stratify the patients in a study and can provide a better characterization of the diseases. Unfortunately, to this regard we note that standard network methods such as Coxnet and ADMMnet do not include procedures for patients stratification in current implementation. This limitation has to be addressed in future works. Third, in order to facilitate the use of the proposed methodology for the analysis of different cancer datasets, it is necessary to implement an interactive user-friendly interface where all preprocessing and normalization steps, as well as the those described in Algorithm 1 can be carried out in terms of a easy point-and-click approach.

AUTHOR CONTRIBUTIONS

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.

FUNDING

This research was supported by POR CAMPANIA FSE 2007/2013—POR CAMPANIA FSE 2014/2020, ASSE IV, CAPITALE UMANO, ASSE V, TRANSNAZIONALITA' ED INTERREGIONALITA' and Italian Flagship Project EPIGEN. This work was supported by GNCS -INDAM.

ACKNOWLEDGMENTS

This study make use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium. Funding for this project was provided by Cancer Research UK and the British Columbia Cancer Agency Branch.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00206/full#supplementary-material

REFERENCES

Ahmad, A., and Fröhlich, H. (2017). Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. *Bioinformatics* 33, 3558–3566. doi: 10.1093/bioinformatics/btx464

Angelini, C., and Costa, V. (2014). Understanding gene regulatory mechanisms by integrating chip-seq and rna-seq data: statistical solutions to biological problems. *Front. Cell Dev. Biol.* 2:51. doi: 10.3389/fcell.2014.00051

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 17:S15. doi: 10.1186/s12859-015-0857-9

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Candes, E., and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.* 35, 2313–2351. doi: 10.1214/009053606000001523

Chang, F., Lee, J., Navolanic, P., Steelman, L., Shelton, J., Blalock, W., et al. (2003). Involvement of pi3k/akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy. *Leukemia* 17, PP590–PP603. doi: 10.1038/sj.leu.2402824

Cohen, L. A., and Guan, J.-L. (2005). Mechanisms of focal adhesion kinase regulation. *Curr. Cancer Drug Targets* 5, 629–643. doi: 10.2174/156800905774932798

Cotterill, S. (1999). *Cancer Genetics Web*. Available online at: http://www.cancer-genetics.org/. (Accessed 21 August, 2015).

Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B* 34, 187–220.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton, FL; London; New York, NY: CRC Press.

Dunning, M., Lynch, A., and Eldridge, M. (2015). *Illuminahumanv4.db: Illumina Humanht12v4 Annotation Data (chip Illuminahumanv4)*. R package version 1.26.0.

Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet* 5:15. doi: 10.3389/fgene.2014.00015

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Am. Stat. Assoc.* 106, 544–557. doi: 10.1198/jasa.2011.tm09779

Fan, J., Feng, Y., and Wu, Y., (2010a). High-dimensional variable selection for cox's proportional hazards model. *Inst. Math. Stat.* 6, 70–86. doi: 10.1214/10-IMSCOLL606

Fan, J., and Li, R. (2001). Variable selection via nonconcave 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. Ser. B* 70, 849–911. doi: 10.1111/j.1467-9868.2008.00674.x

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* 10, 2013–2038. doi: 10.1145/1577069.1755853

Fan, J., and Song, R. (2010b). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Stat.* 38, 3567–3604. doi: 10.1214/10-AOS798

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., et al. (2011). Novel breast cancer susceptibility locus at 9q31. 2: results of a genome-wide association study. *J. Natl. Cancer Inst.* 103, 425–435. doi: 10.1093/jnci/djq563

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., et al. (2010). Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 39, D945–D950. doi: 10.1093/nar/gkq929

Fröhlich, H. (2014). Including network knowledge into cox regression models for biomarker signature discovery. *Biometr. J.* 56, 287–306. doi: 10.1002/bimj.201300035

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.* 6:pl1. doi: 10.1126/scisignal.2004088

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:S3. doi: 10.1186/1752-0509-8-S1-S3

Gremel, G., Ryan, D., Rafferty, M., Lanigan, F., Hegarty, S., Lavelle, M., et al. (2011). Functional and prognostic relevance of the homeobox protein msx2 in malignant melanoma. *Br. J. Cancer* 105:565. doi: 10.1038/bjc.2011.249

Guo, N. L., and Wan, Y.-W. (2014). Network-based identification of biomarkers coexpressed with multiple pathways. *Cancer Inform.* 13, 37–47. doi: 10.4137/CIN.S14054

Gurda, D., Handschuh, L., Kotkowiak, W., and Jakubowski, H. (2015). Homocysteine thiolactone and n-homocysteinylated protein induce pro-atherogenic changes in gene expression in human vascular endothelial cells. *Amino Acids* 47, 1319–1339. doi: 10.1007/s00726-015-1956-7

Hamilton, A., and Piccart, M. (2000). The contribution of molecular markers to the prediction of response in the treatment of breast cancer: a review of the literature on her-2, p53 and bcl-2. *Ann. Oncol.* 11, 647–663. doi: 10.1023/A:1008390429428

Hirata, Y., Kurobe, H., Nishio, C., Tanaka, K., Fukuda, D., Uematsu, E., et al. (2013). Exendin-4, a glucagon-like peptide-1 receptor agonist, attenuates neointimal hyperplasia after vascular injury. *Eur. J. Pharmacol.* 699, 106–111. doi: 10.1016/j.ejphar.2012.11.057

Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G., and Garmire, L. X. (2016). Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome Med.* 8:34. doi: 10.1186/s13073-016-0289-9

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., Occhipinti, A., Angelini, C., De Feis, I., and Lió, P. (2016). Cancer markers selection using network-based cox regression: a methodological and computational practice. *Front. Physiol.* 7:208. doi: 10.3389/fphys.2016.00208

Jaakkola, S., Salmikangas, P., Nylund, S., Lehtovirta, P., Nevanlinna, H., Partanen, J., et al. (1993). Amplification of fgfr4 gene in human breast and gynecological cancers. *Int. J. Cancer* 54, 378–382. doi: 10.1002/ijc.2910540305

Jiang, S., Katayama, H., Wang, J., Li, S. A., Hong, Y., Radvanyi, L., et al. (2010). Estrogen-induced aurora kinase-a (aurka) gene expression is activated by gata-3 in estrogen receptor-positive breast cancer cells. *Hormones Cancer* 1, 11–20. doi: 10.1007/s12672-010-0006-x

Joensuu, H., Pylkkänen, L., and Toikkanen, S. (1994). Bcl-2 protein expression and long-term survival in breast cancer. *Am. J. Pathol.* 145:1191.

Kastan, M. B., and Bartek, J. (2004). Cell-cycle checkpoints and cancer. *Nature* 432:316. doi: 10.1038/nature03097

Kim, E. K., and Choi, E.-J. (2010). Pathological roles of mapk signaling pathways in human diseases. *Biochim. Biophys. Acta* 1802, 396–405. doi: 10.1016/j.bbadis.2009.12.009

Kreisler, A., Strissel, P., Strick, R., Neumann, S., Schumacher, U., and Becker, C. (2010). Regulation of the nrsf/rest gene by methylation and creb affects the cellular phenotype of small-cell lung cancer. *Oncogene* 29:5828. doi: 10.1038/onc.2010.321

Lin, W.-W., and Karin, M. (2007). A cytokine-mediated link between innate immunity, inflammation, and cancer. *J. Clin. Invest.* 117, 1175–1183. doi: 10.1172/JCI31537

Murakami, M., and Hirano, T. (2008). Intracellular zinc homeostasis and zinc signaling. *Cancer Sci.* 99, 1515–1522. doi: 10.1111/j.1349-7006.2008.00854.x

Nass, N., Dittmer, A., Hellwig, V., Lange, T., Beyer, J. M., Leyh, B., et al. (2016). Expression of transmembrane protein 26 (tmem26) in breast cancer and its association with drug response. *Oncotarget* 7:38408. doi: 10.18632/oncotarget.9493

Network, T. C. G. A. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Network, T. C. G. A. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Network, T. C. G. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature* 497, 67–73. doi: 10.1038/nature12113

Parsons, J. T. (2003). Focal adhesion kinase: the first ten years. *J. Cell Sci.* 116, 1409–1416. doi: 10.1242/jcs.00373

Pedley, A. M., and Benkovic, S. J. (2017). A new view into the regulation of purine metabolism: the purinosome. *Trends Biochem. Sci.* 42, 141–154. doi: 10.1016/j.tibs.2016.09.009

Pineda, S., Real, F. X., Kogevinas, M., Carrato, A., Chanock, S. J., Malats, N., et al. (2015). Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. *PLoS Genet.* 11:e1005689. doi: 10.1371/journal.pgen.1005689

Ranzani, M., Iyer, V., Ibarra-Soria, X., Velasco-Herrera, M. D. C., Garnett, M., Logan, D., et al. (2017). Revisiting olfactory receptors as putative drivers of cancer. *Wellcome Open Res.* 2:v1. doi: 10.12688/wellcomeopenres.10646.1

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* 16:85. doi: 10.1038/nrg3868

Sanz, G., Leray, I., Dewaele, A., Sobilo, J., Lerondel, S., Bouet, S., et al. (2014). Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. *PLoS ONE* 9:e85110. doi: 10.1371/journal.pone.0085110

Shannon, P. T., Grimes, M., Kutlu, B., Bot, J. J., and Galas, D. J. (2013). Rcytoscape: tools for exploratory network analysis. *BMC Bioinform.* 14:217. doi: 10.1186/1471-2105-14-217

Shilpi, A., Bi, Y., Jung, S., Patra, S. K., and Davuluri, R. V. (2017). Identification of genetic and epigenetic variants associated with breast cancer

prognosis by integrative bioinformatics analysis. *Cancer Inform.* 16, 1–13. doi: 10.4137/CIN.S39783

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13. doi: 10.18637/jss.v039.i05

Song, R., Lu, W., Ma, S., and Jeng, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* 101, 799–814. doi: 10.1093/biomet/asu047

Sun, H., Lin, W., Feng, R., and Li, H. (2014). Network-regularized high-dimensional cox regression for analysis of genomic data. *Stat. Sin.* 24:1433. doi: 10.5705/ss.2012.317

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Stat. Med.* 16, 385–395.

Tiong, K. H., Tan, B. S., Choo, H. L., Chung, F. F.-L., Hii, L.-W., Tan, S. H., et al. (2016). Fibroblast growth factor receptor 4 (fgfr4) and fibroblast growth factor 19 (fgf19) autocrine enhance breast cancer cells survival. *Oncotarget* 7:57633. doi: 10.18632/oncotarget.9328

van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J., and Wessels, L. F. (2006). Cross-validated cox regression on microarray gene expression data. *Stat. Med.* 25, 3201–3216. doi: 10.1002/sim.2353

Weber, G. (1983). Enzymes of purine metabolism in cancer. *Clin. Biochem.* 16, 57–63. doi: 10.1016/S0009-9120(83)94432-6

Wu, Y. (2012). Elastic net for cox's proportional hazards model with a solution path algorithm. *Stat. Sin.* 22, 270–294. doi: 10.5705/ss.2010.107

Zang, C., Wang, T., Deng, K., Li, B., Hu, S., Qin, Q., et al. (2016). High-dimensional genomic data bias correction and data integration using mancie. *Nat. Commun.* 7:11305. doi: 10.1038/ncomms11305

Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B., and Kuang, R. (2013). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol.* 9:e1002975. doi: 10.1371/journal.pcbi.1002975

Zhao, S. D., and Li, Y. (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *J. Multiv. Anal.* 105, 397–411. doi: 10.1016/j.jmva.2011.08.002

Zhao, X., and Guan, J.-L. (2011). Focal adhesion kinase and its signaling pathways in cell migration and angiogenesis. *Adv. Drug Deliv. Rev.* 63, 610–615. doi: 10.1016/j.addr.2010.11.001

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Am. Stat. Assoc.* 106, 1464–1475. doi: 10.1198/jasa.2011.tm10563

Ziegler, Y. S., Moresco, J. J., Tu, P. G., Yates, J. R. III., and Nardulli, A. M. (2014). Plasma membrane proteomics of human breast cancer cell lines identifies potential targets for breast cancer diagnosis and treatment. *PLoS ONE* 9:e102341. doi: 10.1371/journal.pone.0102341

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101, 1418–1429. doi: 10.1198/016214506000000735

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

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