Lopes *et al. BMC Bioinformatics* (2019) 20: 356
https://doi.org/10.1186/s12859-019-2937-8

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Twiner: correlation-based regularization for identifying common cancer gene signatures

Marta B. Lopes[1,2*], Sandra Casimiro[3] and Susana Vinga[2,4]

Abstract

Background: Breast and prostate cancers are typical examples of hormone-dependent cancers, showing remarkable similarities at the hormone-related signaling pathways level, and exhibiting a high tropism to bone. While the identification of genes playing a specific role in each cancer type brings invaluable insights for gene therapy research by targeting disease-specific cell functions not accounted so far, identifying a common gene signature to breast and prostate cancers could unravel new targets to tackle shared hormone-dependent disease features, like bone relapse. This would potentially allow the development of new targeted therapies directed to genes regulating both cancer types, with a consequent positive impact in cancer management and health economics.

Results: We address the challenge of extracting gene signatures from transcriptomic data of prostate adenocarcinoma (PRAD) and breast invasive carcinoma (BRCA) samples, particularly estrogen positive (ER+), and androgen positive (AR+) triple-negative breast cancer (TNBC), using sparse logistic regression. The introduction of gene network information based on the distances between BRCA and PRAD correlation matrices is investigated, through the proposed *twin networks recovery* (`twiner`) penalty, as a strategy to ensure similarly correlated gene features in two diseases to be less penalized during the feature selection procedure.

Conclusions: Our analysis led to the identification of genes that show a similar correlation pattern in BRCA and PRAD transcriptomic data, and are selected as key players in the classification of breast and prostate samples into ER+ BRCA/AR+ TNBC/PRAD tumor and normal tissues, and also associated with survival time distributions. The results obtained are supported by the literature and are expected to unveil the similarities between the diseases, disclose common disease biomarkers, and help in the definition of new strategies for more effective therapies.

Keywords: Gene network, Sparse logistic regression, Breast invasive carcinoma, Triple-negative breast cancer, Prostate adenocarcinoma

Background

Breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD) are the two most common invasive cancers in women and men, respectively [1]. In both types of cancers, the vast majority of cases are hormone-dependent, meaning that tumor growth is deeply related to hormone-related signaling pathways. About 70% of all BRCA are estrogen receptor (ER) and/or progesterone receptor (PR) positive (ER+ and/or PR+), and endocrine treatment can be effective in all stages of disease [2].

In the PRAD case, androgen/androgen receptor (AR) signaling pathway is deeply involved in the progression of the disease, and androgen deprivation therapy (ADT) with anti-androgens remains as the main treatment in early and late stage disease [3]. Hormone-dependent signaling pathways like ER, PR or AR, ultimately regulate numerous cell functions, and positively impact cell proliferation [4, 5]. However, ER and AR signaling are not exclusively important in BRCA and PRAD, respectively. It is known that estrogens play an important role in male sex hormone secretion, in the physiology of normal prostate

*Correspondence: marta.lopes@tecnico.ulisboa.pt
[1]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
[2]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol 9, 1000-029 Lisboa, Portugal
Full list of author information is available at the end of the article

tissues, and in prostate carcinogenesis. In fact, selective targeting of ERα or β may be an option in the treatment of castration resistant metastatic prostate cancer cells [6]. By the other hand, AR is expressed in about 80% of primary breast cancers, particularly in triple-negative breast cancer (TNBC), characterized by lack of expression of estrogen receptor 1 (*ER*), progesterone receptor (*PR*), and human epidermal growth factor receptor type 2 (*HER2*) [7], and associated with a poor prognosis [8]. AR-inhibiting drugs have indeed shown antitumorigenic activity in preclinical and proof-of-concept clinical studies in TNBC [9].

Given that BRCA and PRAD are hormone-dependent cancers, possibly sharing layers of signaling and regulatory pathways, it is important to unravel common players that could establish a link between the hormone-dependence and the fact that both types of cancers exhibit a high tropism to bone. To achieve that, one possible solution is to, given a classification model, extract the most relevant features in the discrimination between tumor and normal tissue, out of the full set of dozens of thousands of features currently delivered by high-throughtput 'omic technologies. Classification of tumor and normal tissue can be performed separately for BRCA and PRAD, however, given the similarities between the two diseases, a common strategy to classify patients tissue, while simultaneously identifying the genes playing a role in both diseases, would be of great value. This way, a considerable reduction in the effort in defining new therapies could be accomplished.

Statistical learning in high-dimensional 'omic data poses many challenges, in particular for parameter estimation, since the models are seldom identifiable. One way to cope with this problem is to add constraints in the parameters space. For instance, imposing sparsity in the solution will enable feature selection since a large subset of the parameters will be exactly zero. Several methods have been applied in the context of 'omic data, namely the lasso and elastic net, which impose a l_1 regularizer and a linear combination of l_1 (lasso) and l_2 (ridge) penalties, respectively [10, 11]. While the ridge penalty cannot shrink coefficients exactly to zero, therefore keeping all the variables in the model, the lasso estimator enables performing variable shrinkage and selection at the same time, making the solution sparse.

Model constraints can benefit from external knowledge on the biological disease processes, often given by network information. For example, groups of genes are co-expressed under certain conditions or their protein products interact with each other to carry out a biological function [12], which can be represented by graphs. Given a graph $G := (V, E)$, V denotes vertices (or nodes) and E the set of edges. In a gene network, vertices are genes and edges represent a weighted relation between two genes. It has been advocated that incorporating network information as a constraint in the loss function potentially increases the model predictive performance of, e.g., sparse Cox and logistic regression models, as shown when modeling the survival of ovarian cancer carcinoma patients and classifying patients into breast cancer subtypes [12–14]). Furthermore, including network-based regularizers may improve model interpretability since prior knowledge/information via constraints will drive parameter estimation towards meaningful biological solutions. Such network information can be either obtained by a priori defined pathways and network interactions available in public databases, by de novo construction of specific subnetworks from the set of mutated or differentially expressed genes (e.g., [15, 16]), or by the data correlation itself [13].

In this work we combine correlation-based regularizers and sparse logistic regression to solve a binary classification problem with BRCA and PRAD cancer tissues, and normal tissue from breast and prostate cancer patients as classes. Different datasets will be considered in two case studies in the search for shared gene signatures in BRCA and PRAD: I) ER+ BRCA vs. PRAD, showing similarities at the ER signaling level and shared marked bone osteotropism; and II) AR+ TNBC vs. PRAD, sharing AR-signaling dependency. With the goal of identifying a common network to each BRCA subtype and PRAD data, ER+ BRCA, AR+ TNBC and PRAD gene correlation networks will be generated using the Pearson correlation between observed (gene expression) variables as similarity measure. For a given gene in the network, the more similar its correlation pattern between the two diseases under consideration is, the less penalized it will be in the regularization term of sparse logistic regression. Selected similarly correlated genes in the two diseases, i.e., playing a role in discriminating between cancer and non-cancer, can be seen as potential biomarkers candidates in the two groups of patients.

Methods

Datasets

The transcriptomic data on breast and prostate cancer patients used in this work were obtained from The Cancer Genome Atlas (TCGA) Data Portal (https://cancergenome.nih.gov/).

Breast invasive carcinoma (BRCA)

The BRCA RNA-Seq Fragments Per Kilo base per Million (FPKM) dataset was imported using the 'brca.data' R package[1]. The BRCA gene expression data is composed of 57251 variables for a total of 1222 samples from 1097 individuals. From those samples, 1102 correspond to primary solid tumor, 7 to metastases and 113 to normal breast tissue. Only samples from primary solid tumor were selected for analysis, from those only ER+ BRCA

samples were considered to avoid the introduction of confounding effects by accounting for non-hormonal BRCA tumor information in the search for a common gene signature for BRCA and PRAD. Information on the samples 'positive' clinical status for ER was obtained from the BRCA clinical data also available from the TCGA. The BRCA response variable $\mathbf{Y}$ is binary, coded with '1' for *tumor* (802 samples) and '0' for *normal* (79 samples) tissue.

Triple-negative breast cancer (TNBC)

The TNBC dataset was built from the BRCA dataset described above. The TNBC binary response vector $\mathbf{Y}$ was created, with '1' corresponding to TNBC individuals (with *ESR1*, *PGR* and *ERBB2* 'negative' expression), and non-TNBC ('0') to non-TNBC (other types of BRCA) patients, whenever at least one of the three genes is 'positive'. The individuals' *status* regarding ER, PR and HER2, needed for building $\mathbf{Y}$, were obtained from the BRCA clinical data available from the TCGA, composed of 114 variables, as described in Lopes et al. (2018) [17, 18]. Only AR+ TNBC samples (with AR expression larger than the median AR expression over all TNBC samples) were considered for building the TNBC dataset, accounting for 80 *tumor* and 113 *normal* tissue samples.

Prostate adenocarcinoma (PRAD)

The PRAD RNA-Seq Fragments Per Kilo base per Million (FPKM) dataset was imported using the 'prad.data' R package[2]. The PRAD gene expression data is composed of 57035 variables for a total of 551 samples from 500 individuals. From those samples, 495 correspond to primary solid tumors, 1 to metastases and 52 to normal tissue. Only samples from primary solid tumor were considered for analysis. The PRAD response variable $\mathbf{Y}$ is binary, coded with '1' for *tumor* samples and '0' for *normal* tissue samples. The PRAD dataset is composed of 495 *tumor* and 52 *normal* samples.

Data pre-processing

FPKM normalized ER+ BRCA, AR+ TNBC, and PRAD gene expression data were log-transformed and Z-score normalized prior to data analysis. A subset of $\sim$20000 variables in each dataset was considered for further analysis, corresponding to the protein coding genes reported from the Ensembl genome browser [19] and the Consensus CDS project [20], and shared by each pair of diseases under evaluation (ER+ BRCA vs. PRAD and AR+ TNBC vs. PRAD).

Classification modeling

Sparse logistic regression

Binary logistic regression describes the relationship between one or more independent variables and a binary outcome vector $\mathbf{Y} = \{Y_i\}_{i=1,...,n}$, which is given by the logistic function

$$P(Y_i = 1|\mathbf{X}_i) = \frac{\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}, \tag{1}$$

where $\mathbf{X}_i, i = 1, ..., n$, is the vector of p covariates for observation i, $P(Y_i = 1|\mathbf{X}_i)$ is the probability of success for observation i, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_p)$ are the regression coefficients associated with the p independent variables.

The parameters of the logistic model are estimated by maximizing the log-likelihood function of the logistic model, given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{y_i \log P(Y_i = 1|\mathbf{X}_i) + (1-y_i)\log\left[1 - P(Y_i = 1|\mathbf{X}_i)\right]\right\}, \tag{2}$$

where the binary variable y_i indicates success ($y_i = 1$) or unsuccess ($y_i = 0$) for observation i. By the introduction of a regularization term, the log-likelihood function becomes

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{y_i \log P(Y_i = 1|\mathbf{X}_i) + (1-y_i)\log\left[1 - P(Y_i = 1|\mathbf{X}_i)\right]\right\} + F(\boldsymbol{\beta}), \tag{3}$$

where

$$F(\boldsymbol{\beta}) = \lambda\left\{\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2\right\} \tag{4}$$

stands for the elastic net penalty, with $\alpha = 1$ corresponding to lasso and $\alpha = 0$ to ridge, and the tuning parameter λ controlling the amount of shrinkage in the coefficients.

Correlation-based network regularization

With the specific goal of weighting variables based on their similarities across two given diseases, we propose `twiner`, a structured regularizer based on the pairwise correlations between variables, independently obtained from two given datasets.

Consider two correlation matrices for diseases A and B, $\Sigma_A = \left[\boldsymbol{\sigma}_1^A, ..., \boldsymbol{\sigma}_p^A\right]$ and $\Sigma_B = \left[\boldsymbol{\sigma}_1^B, ..., \boldsymbol{\sigma}_p^B\right]$, respectively, where each column $\boldsymbol{\sigma}_j \in \mathbb{R}^p$ represents the correlation of each gene $j = 1, \ldots, p$ with the remaining ones. The proposed dissimilarity measure $d_j(A, B)$ of gene j between A and B is given by the angle of the corresponding vectors, i.e.,

$$d_j(A,B) = \arccos\frac{<\boldsymbol{\sigma}_j^A, \boldsymbol{\sigma}_j^B>}{\|\boldsymbol{\sigma}_j^A\| \cdot \|\boldsymbol{\sigma}_j^B\|}, \quad j = 1, \ldots, p. \tag{5}$$

The rationale of using the angle is that two patterns will be identified as similar if they have the same proportionality between the entries across the two datasets, irrespective of the magnitude of the vectors. In the context of the present application, one gene has a similar

role in BRCA and PRAD, if it is similarly correlated with the remaining genes in the two diseases. This correlation-based regularization constitutes the basis of `twiner`, since this dissimilarity will then be used as a penalization for the cost function. The weighting vector to be used as a penalization $\mathbf{w} = (w_1, ..., w_j, ..., w_p)$ is therefore based on this distance, normalized by their maximum value:

$$w_j = \frac{d_j(A,B)}{\max_k d_k(A,B)}, \quad j,k = 1,\ldots,p. \tag{6}$$

The penalty term in Eq. 4 takes the form

$$F(\boldsymbol{\beta}) = \lambda \left\{ \alpha \|\mathbf{w} \circ \boldsymbol{\beta}\|_1 + (1-\alpha)\|\mathbf{w} \circ \boldsymbol{\beta}\|_2^2 \right\}, \tag{7}$$

with vector $\mathbf{w}$ representing the factors that control how much of the penalty λ affects each coefficient, and $\circ$ standing for the element wise (or Hadamard) product.

For a given gene *j*, the smaller the distance between $\boldsymbol{\sigma}_j^A$ and $\boldsymbol{\sigma}_j^B$, the more similar the diseases are regarding the overall gene *j* correlation pattern. Therefore, the resulting *twin networks recovery* (`twiner`) penalty enables the identification of variables with similar (*twin*) correlation with the remaining variables across the two diseases, with smaller penalties being associated to genes with smaller distances between the diseases' correlation matrices. The influence these genes have in the outcome will be assessed by regularized logistic regression based on the elastic net penalty.

The modeling strategy described above will be applied in two case studies with the goal of finding common gene signatures in i) ER+ BRCA and PRAD, and ii) AR+ TNBC and PRAD, as described next.

Classification of BRCA and PRAD RNA-Seq data

Sparse logistic regression using the elastic net penalty (EN) was used to classify RNA-Seq data from patients into ER+ BRCA, AR+ TNBC and PRAD vs. *normal* breast and prostate tissue samples. The overall procedure for dataset construction is illustrated in Fig. 1.

With the goal of finding a common gene signature between ER+ BRCA and PRAD cancer types, ER+ BRCA and PRAD data were grouped into a single class, i.e., *tumor*, and EN applied to classify RNA-Seq data into ER+ BRCA or PRAD (*tumor*) vs. *normal* samples, herein called BRCAPRAD model. Three quarters of randomly selected samples were assigned to training samples for model construction, whereas the remaining samples were assigned to test samples for model evaluation. The classification was performed using two models: 1) EN; and 2) sparse logistic regression using the `twiner` penalty (twiner). For both EN and `twiner` models the alpha parameter was set to $\alpha = 0.9$, which yields a adequate number of features to be further analysed without compromising clinical interpretability. The Pearson correlation matrices from ER+ BRCA and PRAD RNA-Seq data are matrices Σ_A and Σ_B, as seen above, and the response $\mathbf{Y}$ vector is a binary vector with '0' corresponding to *normal* tissue and '1' to *tumor* (ER+ BRCA or PRAD) tissue. The angular distance between $\boldsymbol{\sigma}_j^A$ and $\boldsymbol{\sigma}_j^B$, corresponding to the correlation pattern of variable *j* in matrices Σ_A and Σ_B was used for building the weight vector $\mathbf{w}$ as explained above. With the goal of searching for shared disease biomarkers, genes showing larger angular distances between correlation vectors in the two diseases were discarded, only keeping those (out of the previous $\sim$ 20000) showing an angular distance less than 75°, at the same time contributing to reduce model complexity. The new dimension in the variables space was thus decreased to 16367 genes.

EN and `twiner` models were generated 100 times for randomly chosen training and test sets. The median values for the mean squared error (MSE) of classification, the area under the Precision-Recall curve (AUC) [21, 22] and the number of misclassifications, along with the set of variables selected in more than 75% of the runs, were taken for comparison of the modeling strategies employed.

The same analysis was performed in the search for a common gene signature between AR+ TNBC and PRAD cancer types. As for the ER+ BRCA vs. PRAD case, only genes (out of the previous $\sim$ 20000) showing lower angular distances were considered, yielding 14598 genes for model building in the AR+ TNBC vs. PRAD case. EN and `twiner` models were generated to classify RNA-Seq data into AR+ TNBC or PRAD *tumor* samples ('1') vs. TNBC and PRAD *normal* samples ('0'), herein called the TNBCPRAD model.

For biological interpretation of the variables selected by the methods, gene correlation networks were represented only for ER+ BRCA, AR+ TNBC, and PRAD data, using the variables exclusively selected by EN, `twiner` and shared by the two models.

Finally, as an attempt to clinically validate our approach, the variables selected by EN and `twiner` were tested in a survival analysis on ER+ BRCA, AR+ TNBC and PRAD tumor data, using the Cox regression model [23]. The individuals in each dataset were separated in two groups by the median of the fitted relative risk. This allows to perform the Log-Rank test via the Kaplan-Meier estimator [24], and to assess if the survival curves of the two groups are statistically different by calculating the *p*-values. Increased risk groups' separability by a given set of genes is expected to trigger further research on the role of these genes in the disease.

Individual models for predicting the class, *tumor* vs. *normal*, independently for ER+ BRCA, AR+ TNBC and PRAD data, herein called ER+ BRCA, AR+ TNBC and PRAD models, were also built, using the same set

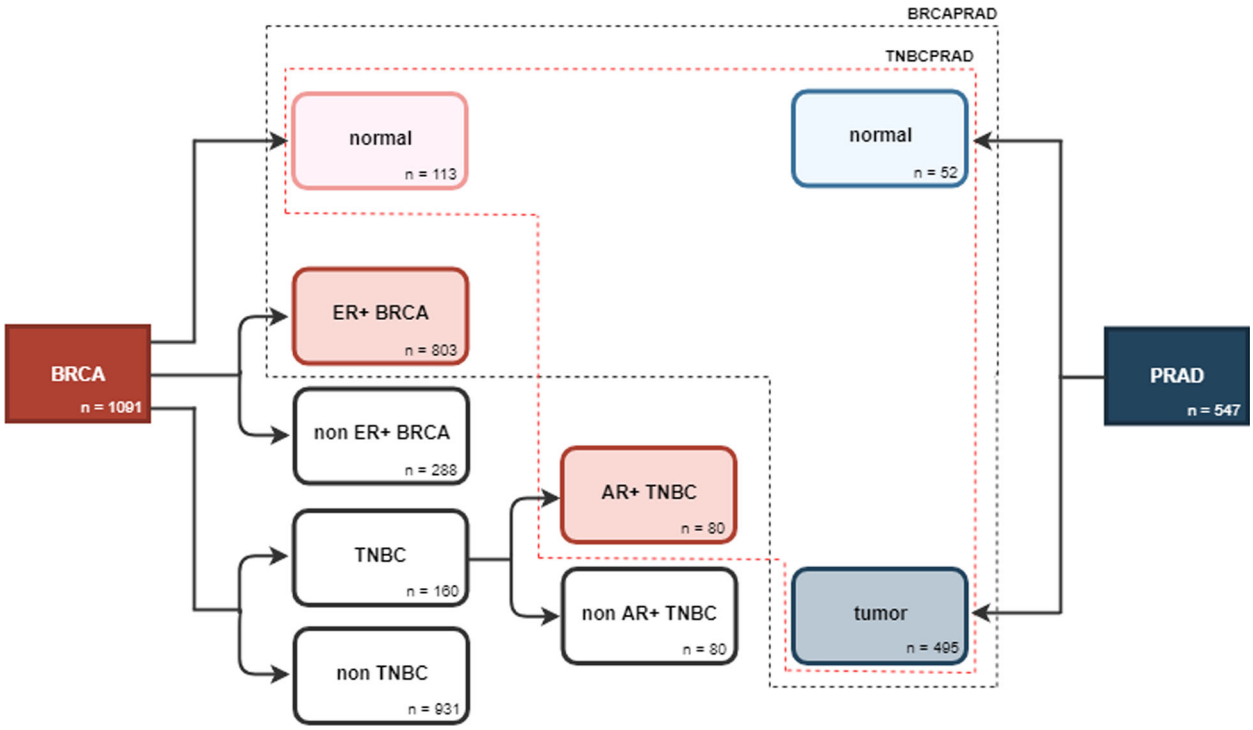

Fig. 1 Schematic representation of the procedure to construct the BRCAPRAD (blue dashed line) and TNBCPRAD (red dashed line) datasets analysed. BRCA, Breast Invasive Carcinoma; TNBC, Triple-Negative Breast Cancer; ER+ BRCA, estrogen receptor positive BRCA; AR+ TNBC, androgen receptor positive TNBC; PRAD, Prostate Adenocarcinoma

of variables used for the combined BRCAPRAD and TNBCPRAD models, as explained above. The goal was to identify (if any) genes selected in common by independent disease models (potential shared disease biomarkers), and the overlap with BRCAPRAD and TNBCPRAD sparse logistic models aiming at extracting common gene signatures from two diseases through the `twiner` penalty. Similarly to the independent BRCAPRAD nd TNBCPRAD models, training and test sets were randomly generated. The optimization of the parameters λ and α based on the MSE for the models described above was performed by 10-fold cross-validation (CV), with varying α values ($1 > \alpha > 0$) tested.

The `glmnet` R package [25] implemented in the free R statistical software [26] was used in our study for building the above sparse logistic regression models with elastic net regularization. The **w** vector was introduced as penalty factor in the `glmnet` function. Differentially expressed genes across *tumor* and *normal* ER+ BRCA, AR+ TNBC and PRAD tissue were identified using the `limma` Bioconductor R package [27], in order to support further clinical analysis and interpretation of the obtained genes.

Results

Principal component analysis

Before classification of samples into ER+ BRCA/AR+ PRAD and PRAD *tumor* and *normal* breast and prostate samples, a first non-supervised analysis was intended to visualize samples' grouping in a reduced dimensional space. A Principal Component Analysis (PCA) was applied to a dataset comprising gene expression data from ER+ BRCA, AR+ TNBC and PRAD *tumor* tissue samples, along *normal* tissue samples from breast and prostate patients. A clear separation between ER+ BRCA/AR+ TNBC and PRAD *tumor* samples is observed in the space of the first two principal components (Fig. 2), though partial overlap is observed when looking at PCs individually. Overlap between ER+ BRCA/AR+ TNBC *tumor* and *normal* samples is absent in PC2, as opposed to that observed for PRAD *tumor* and *normal* samples, showing overlap in both PCs. Finally, great overlap between BRCA (ER+ BRCA) and TNBC (AR+ TNBC) is observed in the subspace represented (Fig. 2).

Sparse logistic regression

Sparse logistic regression models were built independently for ER+ BRCA, AR+ TNBC and PRAD data for the classification of patients into *tumor* or *normal* tissue. The ER+ BRCA model was based on $\alpha = 0.9$, ending up in the selection of 42 variables and 1 misclassification in the test set (Table 1). The model generated for AR+ TNBC dataset was based on an α value of 0.8, yielding the selection of 63 variables and no misclassifications in both training and

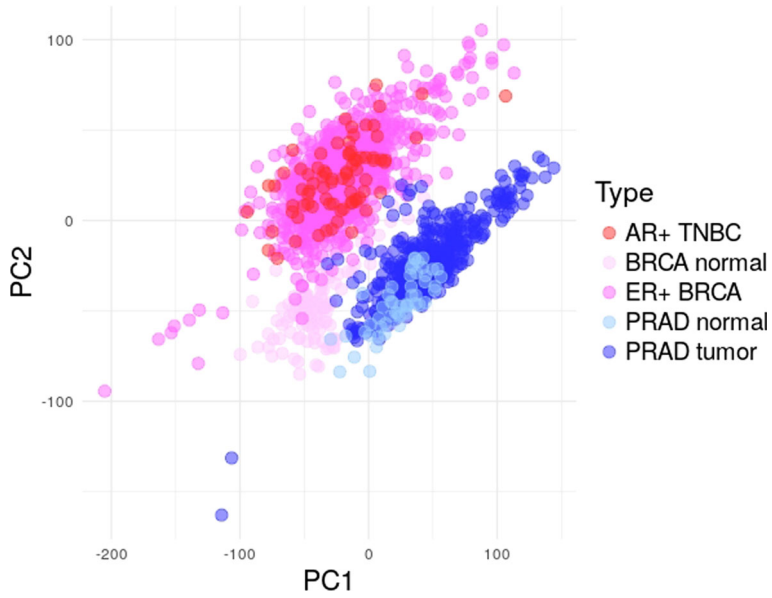

Fig. 2 Representation of ER+ BRCA, AR+ TNBC and PRAD tumor and normal samples in the space of the first two principal components. ER+ BRCA, estrogen receptor positive Breast Invasive Carcinoma; AR+ TNBC, androgen receptor positive Triple-Negative Breast Cancer; PRAD, Prostate Adenocarcinoma

test sets. The PRAD model selected 68 variables, considering an optimum α value of 0.5, and misclassified 11 patients in the training set and 6 in the test set. A higher number of misclassifications was obtained for the PRAD dataset, as foreseen by PCA (Fig. 2). No variables were selected in common between the ER+ BRCA and PRAD models (Fig. 3a), whereas only one variable, gene *NKAPL*, was selected in common between the TNBC and PRAD models using this strategy (Fig. 3b).

ER+ BRCA vs. PRAD

EN and `twiner` were applied to the BRCAPRAD dataset, as a means to identify a common gene signature between ER+ BRCA and PRAD diseases. Summary results of the two modeling strategies applied to 100 random training and test sets can be found in Table 2. A median MSE decrease in 11% and 4% for the training and test sets was observed, respectively (Table 2). Three genes were always selected by the 100 EN models (*BGN*, *GLRA4* and *NKAPL*) and 2 by `twiner` models (*NKAPL* and *PAK3*), with *NKAPL* being selected in common by the two modelling strategies.

Seventeen genes were selected in more than 75 out of 100 `twiner` models (Fig. 3; Table 3), 8 in common with EN (*CXCR2*, *GLRA4*, *LRRC3B*, *NKAPL*, *PAK3*, *RP11-729L2.2*, *SCN5A* and *TMEM236*) and the remaining 9 (*BMT2*, *CLEC11A*, *CSGALNACT2*, *HMGCS1*, *POLR2H*, *RP11-371E8.4*, *SCARA5*, *SLC17A7* and *ZBTB24*) exclusively selected by `twiner` (Fig. 3; Table 3). Out of the genes selected by `twiner`, 4 (*GLRA4*, *LRRC3B*, *PAK3* and *SLC17A7*) were shared with the BRCA model and 1 (*NKAPL*) with the PRAD model, respectively (Table 3). Most genes from the 9 genes exclusively selected by `twiner` (Fig. 3a) have lower weights compared to those exclusively selected by EN and selected in common by the two strategies (Fig. 4a), meaning that their correlation pattern across the genes space is more similar between ER+ BRCA and PRAD cancer types, compared to the remaining genes selected.

Table 1 Summary of BRCA, TNBC and PRAD EN models (MSE, mean squared error; AUC, area under the precision-recall curve; Miscl, misclassifications; Vars, nr. of variables selected)

	α	# Vars	MSE		AUC		# Miscl	
			Train	Test	Train	Test	Train	Test
BRCA	0.9	42	0.0001	0.0065	1	1	0	1
TNBC	0.8	63	0.0002	0.0025	1	1	0	0
PRAD	0.5	68	0.0187	0.0309	0.97	0.97	11	6

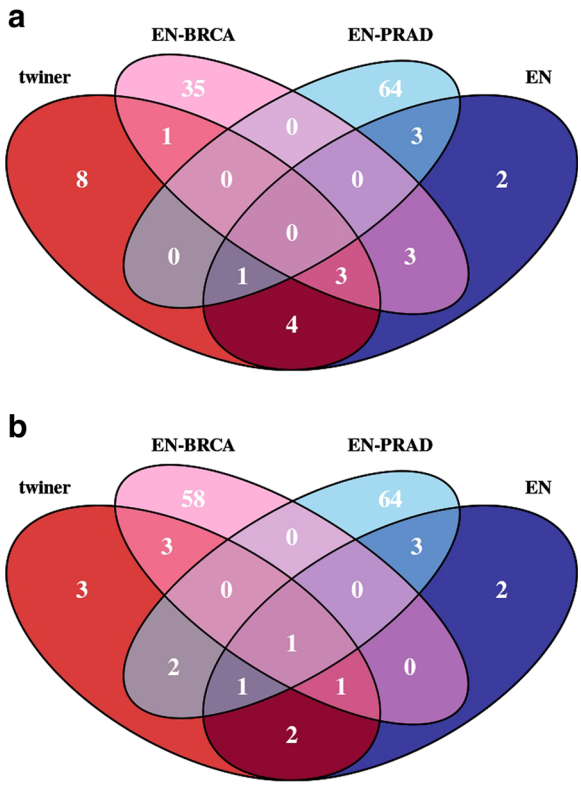

Fig. 3 Venn diagrams representing the number of variables selected by elastic net (EN) (blue) and `twiner` (red), and by EN-BRCA (pink) and EN-PRAD (light blue) individual models, for the two case studies evaluated: **a**) ER+ BRCA vs. PRAD; and **b**) AR+ TNBC vs. PRAD. ER+ BRCA, estrogen receptor positive Breast Invasive Carcinoma; AR+ TNBC, androgen receptor positive Triple-Negative Breast Cancer; PRAD, Prostate Adenocarcinoma

Figure 5 shows the correlation gene networks for the ER+ BRCA and PRAD *tumor* and *normal* samples regarding the genes exclusively selected in more than 75 BRCAPRAD EN and `twiner` models built on 100 randomly selected training and test sets, as well as the genes selected in common by the two model strategies. The thickness of the connecting lines represents the strength of the correlation, whereas green represents a positive correlation and red signals a negative correlation. In the search for shared disease biomarkers, and towards laboratory and clinical validation, particular attention might be given to the similarities between the relationships across less penalized selected genes (red coloured genes) and common (green) genes in ER+ BRCA and PRAD tumor samples, compared to that observed for BRCA and PRAD normal tissue samples.

AR+ TNBC vs. PRAD

With the goal of finding a common gene signature in AR+ TNBC and PRAD, EN and `twiner` were applied to the TNBCPRAD dataset. A less pronounced model accuracy improvement was obtained for `twiner` over EN, compared to the BRCAPRAD case, with a median MSE decrease in 5% for the training set (Table 2).

Table 2 Summary of BRCAPRAD and TNBCPRAD model results by EN and `twiner`, considering the median values obtained from 100 models built on randomly generated training and test sets (MSE, mean squared error; AUC, area under the precision-recall curve; Miscl, misclassifications; Vars, nr. of variables selected)

		# Vars	MSE		AUC		# Miscl	
			Train	Test	Train	Test	Train	Test
BRCAPRAD	EN	58	0.013	0.018	0.98	0.98	17	9
	`twiner`	69	0.012	0.018	0.99	0.98	15	9
TNBCPRAD	EN	61	0.025	0.034	0.97	0.96	16	8
	`twiner`	71	0.025	0.034	0.97	0.96	14	8

Table 3 Genes selected by EN and `twiner`; pink and blue arrows indicate up- (↑) and down-regulated (↓) genes in ER+ BRCA, AR+ TNBC and PRAD, respectively

ER+ BRCA vs. PRAD				
Exclusively selected by EN				
↑↑*BGN*	↓*DNAJB1*	↓↓*DUOXA2*	↓↓*HSD17B13*	↓↓*KCNS1*
↓↓*KY*	↑↑*RAB17*	↓↓*RP11-903H12.5*		
Exclusively selected by `twiner`				
↓↓*BMT2*	↑↑*CLEC11A*	*CSGALNACT2*	↑↓*HMGCS1*	↑↑*POLR2H*
↑*RP11-371E8.4*	↓↓*SCARA5*	↓ *SLC17A7*	↑↑*ZBTB24*	
Shared between EN and `twiner`				
↓↓*CXCR2*	↓↓*GLRA4*	↓↓*LRRC3B*	↓↓*NKAPL*	↓↓*PAK3*
↓↓*RP11-729L2.2*	↓↓*SCN5A*	↓↓*TMEM236*		
Shared between `twiner` and EN-BRCA				
↓↓*GLRA4*	↓↓*LRRC3B*	↓↓*PAK3*	↓ *SLC17A7*	
Shared between `twiner` and EN-PRAD				
↓↓*NKAPL*				
AR+ TNBC vs. PRAD				
Exclusively selected by EN				
↑↑*C3orf80*	↓↓*CXCR2*	↑↑*LEMD2*	↓↓*SDHD*	↑↑*SIM2*
Exclusively selected by `twiner`				
↓↓*CD300LG*	↑↑*CTU1*	↓↓*KLHL4*	↓↓*PARK2*	↓↓*SCARA5*
↓↑*SLC35E2*	↓↓*SNCG*	↑↑*UCN*		
Shared between EN and `twiner`				
↑↑*BGN*	↓*DNAJB1*	↓↓*GLRA4*	↓↓*GSTM3*	↓↓*NKAPL*
Shared between `twiner` and EN-TNBC				
↑↑*BGN*	↓↓*CD300LG*	↓↓*NKAPL*	↓↓*PARK2*	↓↓*SCARA5*
Shared between `twiner` and EN-PRAD				
↑↑*CTU1*	↓*DNAJB1*	↓↓*KLHL4*	↓↓*NKAPL*	

Thirteen genes were selected in more than 75 out of 100 `twiner` models (Fig. 3; Table 3), 5 in common with EN (*BGN*, *DNAJB1*, *GLRA4*, *GSTM3* and *NKAPL*), and the remaining 8 (*CD300LG*, *CTU1*, *KLHL4*, *PARK2*, *SCARA5*, *SLC35E2*, *SNCG* and *UCN*) exclusively selected by `twiner` (Fig. 3b; Table 3). A total of 5 genes were selected in common between `twiner` and the individual AR+ TNBC model (*BGN*, *CD300LG*, *NKAPL*, *PARK2* and *SCARA5*) and 4 (*CTU1*, *DNAJB1*, *KLHL4* and *NKAPL*) with the PRAD model (Table 3). From the 8 genes exclusively selected by `twiner` (Fig. 3b), particularly 2 (*SLC35E2* and *UCN*) have lower weights compared to the remaining genes exclusively selected by EN and selected in common by the two strategies (Table 3; Fig. 4b), corresponding to two genes that show a similar correlation pattern across AR+ TNBC and PRAD cancer types, and that are relevant in the classification of breast/prostate tissue into *tumor* (AR+ TNBC/PRAD) and *normal* tissue.

As for the ER+ BRCA vs. PRAD case, the correlation networks for the genes selected were obtained for the AR+ TNBC vs. PRAD *tumor* and *normal* samples (Fig. 6). In both case studies, the relationships highlighted in the correlation networks for the relevant genes selected by our analysis is expected to be matter for further disease understanding and biomarker research.

Survival analysis

With the goal of assessing the clinical significance of accounting for the variables selected by our method, the genes selected by EN and `twiner` were tested in a survival analysis using the Cox regression model on ER+ BRCA, AR+ TNBC and PRAD tumor data. This constitutes an external and independent evaluation of the genes selected under the context of survival analysis, with the goal of expanding the usefulness of this approach to different clinical data types. Indeed, if the

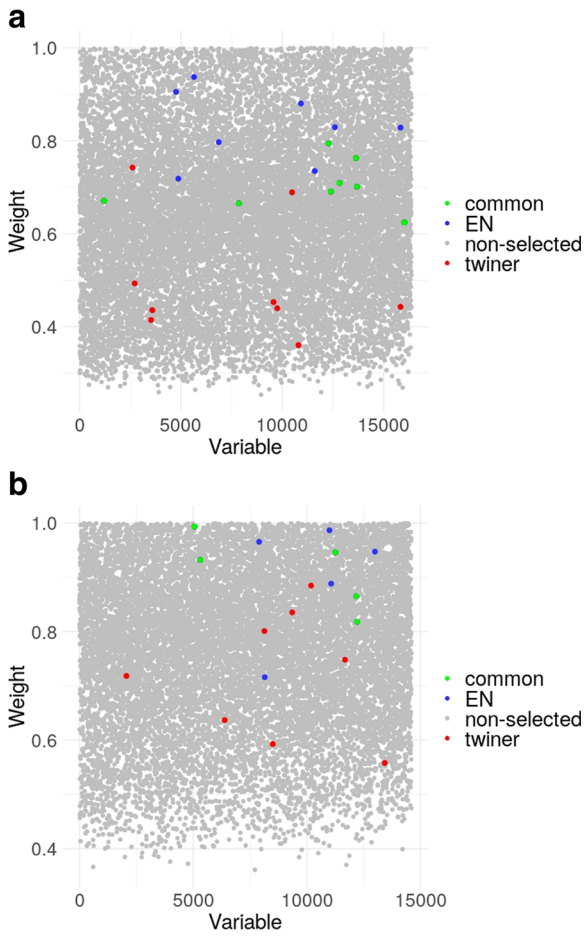

Fig. 4 Weights of the variables selected by elatic net (EN) (blue) and `twiner` (red), and selected in common by the two modeling strategies (green), for the two case studies evaluated: **a**) ER+ BRCA vs. PRAD; and **b**) AR+ TNBC vs. PRAD. ER+ BRCA, estrogen receptor positive Breast Invasive Carcinoma; AR+ TNBC, androgen receptor positive Triple-Negative Breast Cancer; PRAD, Prostate Adenocarcinoma

common gene signatures identified are also associated with the distribution of follow-up times and status, this would strengthen correlation-based regularization as a promising method to support prognostic assessment in cancer studies.

Figures 7 and 8 show the survival curves obtained for each dataset considered in the ER+ BRCA vs. PRAD and AR+ TNBC vs. PRAD cases, respectively. In the first case (Fig. 7), significance by the log-rank test for the difference in the survival distributions for high and low risk patients (separated by the median of the fitted relative risk) is highly increased in ER+ BRCA, while for PRAD the difference becomes significant. In the second case (Fig. 8), the difference for high and low risk individuals becomes statistically significant in both AR+ TNBC and PRAD. These results clearly indicate that accounting for genes showing a similar correlation pattern across the diseases, and without losing predictive ability, indeed improves the separation of high and low-risk patients.

Discussion

After analyzing the computational results, an evaluation of the role of the genes selected in the diseases studied becomes crucial when the goal is to unravel new targets to tackle shared disease features. A discussion on the relevance of the selected genes on the clinical and oncobiology of breast and prostate cancers can be found next.

ER+ BRCA is the most frequent subtype of breast cancer amongst women. ER+ BRCA shares with PRAD the response to hormone signaling and marked bone osteotropism. Bone metastases are the principal site of metastasis in ER+ BRCA and PRAD, significantly affecting morbidity and mortality. However, it is still unknown if and how hormone signaling is involved in specific biological features of bone tropic metastasis initiating cells.

In this work we proposed to identify events specifically deregulated in both ER+ BRCA and PRAD cancers. We hypothesized that these events could reflect a possible

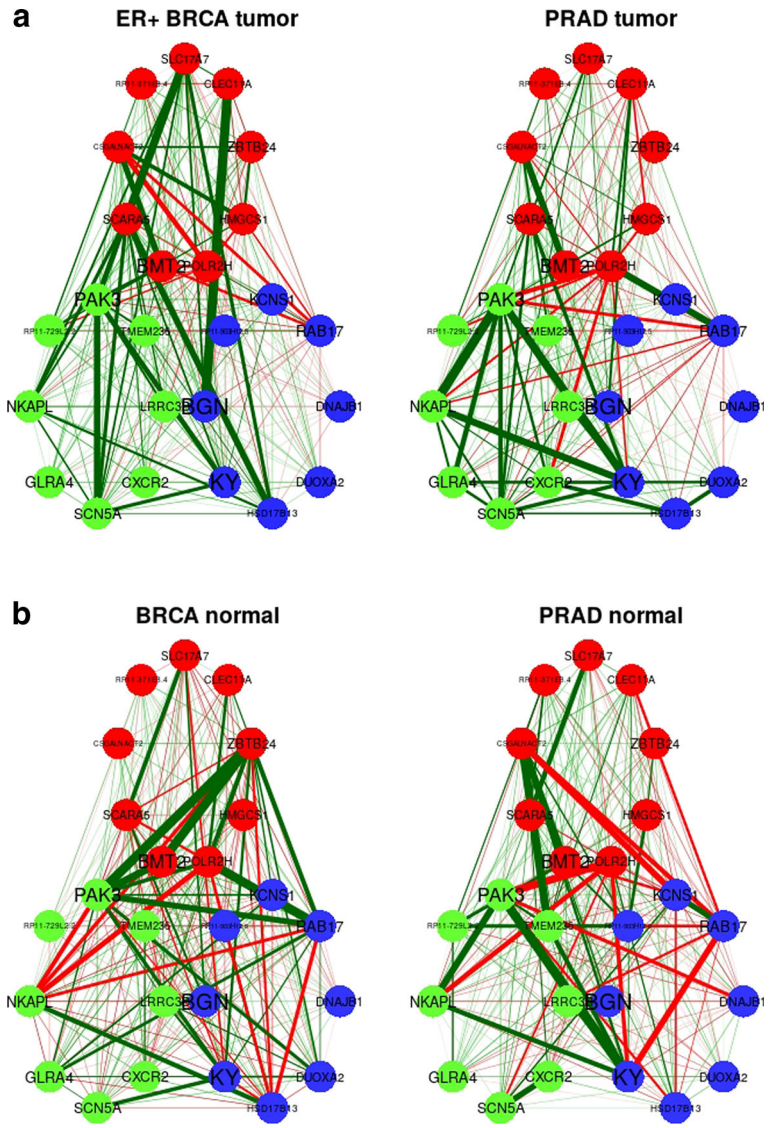

Fig. 5 Network representation of the correlation between the genes selected by elastic net (EN) and `twiner` in the ER+ BRCA vs. PRAD case study; **a**) and **b**) stand for *tumor* and *normal* samples, respectively. ER+ BRCA, estrogen receptor positive Breast Invasive Carcinoma; PRAD, Prostate Adenocarcinoma

association with bone tropism and/or hormone signaling. Our analysis led to the identification of nine genes exclusively selected by `twiner`, five identically deregulated in ER+ BRCA and PRAD (Table 3). Amongst these, three genes were up-regulated (*CLEC11A*, *POLR2H* and *ZBTB24*) and two down-regulated (*BMT2* and *SCARA5*). Amongst the three up-regulated, *CLEC11A* may be directly implicated in bone tropism or bone metastasis development.

CLEC11A was previously found to be part of a gene set up-regulated in cancer stem cell populations upon therapeutic insult [28]. *CLEC11A* encodes a C-type lectin domain protein, osteolectin, which is an osteogenesis driver usually secreted by bone stromal cells that promotes the differentiation of mesenchymal progenitors into mature osteoblasts [29]. Therefore, identification of *CLEC11A* in our model may be associated with a blastic bone metastasis phenotype, typical in PRAD and also frequent in BRCA.

In silico analysis had previously identified *POLR2H* as one of the key genes involved in the occurrence of PRAD, and *POLR2H* protein was significantly upregulated in PRAD tissues [30]. However, its role in cancer needs to be further explored.

Finally, *ZBTB24* encodes the poorly characterized zinc finger and BTB domain containing 24 protein, which belongs to the large ZBTB family of transcriptional repressors. Although its function is still unknown, it was recently

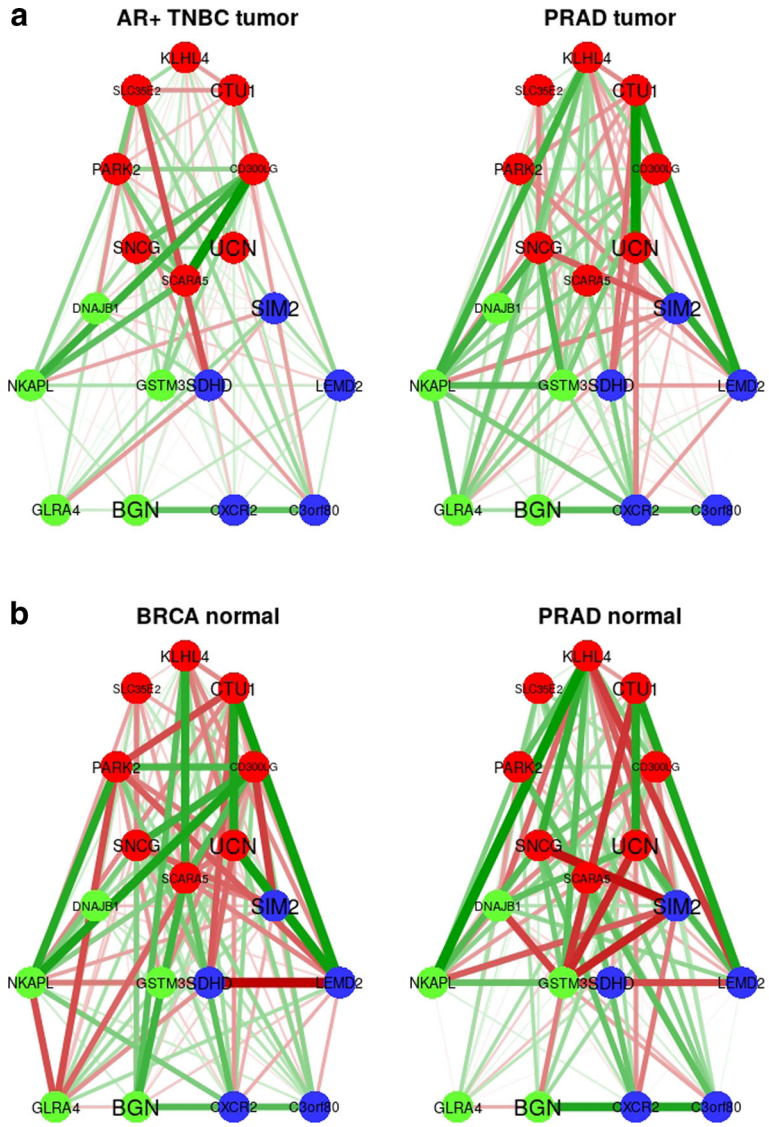

Fig. 6 Network representation of the correlation between the genes selected by elastic net (EN) and `twiner` in the AR+ TNBC vs. PRAD case study; **a**) and **b**) stand for *tumor* and *normal* samples, respectively. AR+ TNBC, androgen receptor positive Triple-Negative Breast Cancer; PRAD, Prostate Adenocarcinoma

shown that *ZBTB24* is involved in the control of DNA methylation [31]. It will be important to address its specific role in BRCA and PRAD.

Amongst the down-regulated genes, *SCARA5* (Scavenger receptor class A member 5) is a candidate tumor suppressor in several malignancies; however, its role in BRCA cell growth and metastasis is still unclear. *SCARA5* was found to be down-regulated in BRCA tissues and cells and correlated with clinicopathologic characteristics [32]. In this study, *SCARA5* overexpression significantly suppressed cell proliferation, colony formation, invasion, and migration, and induced G0/G1 arrest and apoptosis. Recently, another group reported that *SCARA5* expression was significantly decreased in tumors (92.2%), compared to non-cancerous tissue samples, due to the hypermethylation of the promoter [33]. There are no reports implicating *SCARA5* in PRAD, but based on our results we hypothesize a similar pattern of expression. *BMT2* has not been previously implicated in BRCA or PRAD.

AR (androgen receptor)-signaling is particularly important in prostate cancer, however, AR is also expressed in up to 90% of ER+ BRCA, and to a lesser degree, in *HER2* amplified tumors [34]. Although in BRCA HER2+ AR does not seem to play a role, in ER+ BRCA, AR signaling has been correlated with a better prognosis due to its inhibitory activity, but it also may increase resistance to anti-estrogen therapies such as tamoxifen. AR blockade

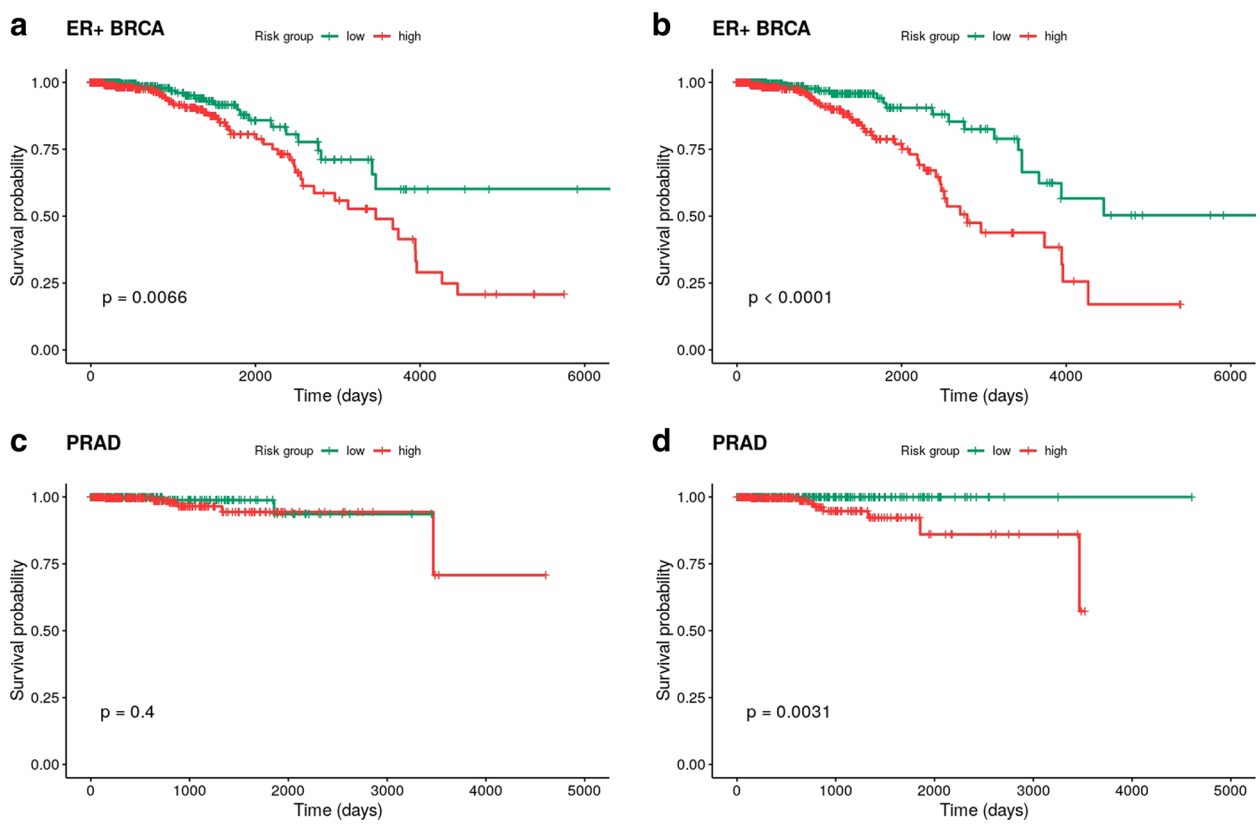

Fig. 7 Kaplan-Meier survival curves obtained for the ER+ BRCA and PRAD datasets using the variables selected by elastic net (EN) (**a** and **c**) and `twiner` (**b** and **d**). ER+ BRCA, estrogen receptor positive Breast Invasive Carcinoma; PRAD, Prostate Adenocarcinoma

can resensitize cells, and therefore is potential target in ER+ breast cancer. In TNBC, gene expression profiling studies have led to the identification of a luminal androgen receptor (LAR) subtype that is dependent on AR signaling, and there seems to be an association between AR expression and improved outcomes in TNBC. Clinical studies targeting AR have indeed shown promising results in this setting. Although to a significant less extent, TNBC may also metastize to the bone, however, the incidence of bone metastasis is significantly higher for the LAR subtype [35]. Therefore, we interrogated if we also could find common genes deregulated in TNBC with elevated AR expression, AR+ (AR values > median AR expression) and PRAD, AR-dependent. We found seven genes equally deregulated in both AR+ TNBC and PRAD (Table 3): *CTU* and *UCN*, up-regulated; and *CD300LG*, *KLHL4*, *PARK2*, *SCARA5* and *SNCG*, down-regulated. Only *SCARA5* was detected in the ER+ BRCA vs. PRAD analysis, suggesting specificity for AR+ TNBC.

CTU1 (cytosolic tRNA thiouridilase) is involved in maintaining genome stability, since post-transcriptional modifications of transfer RNAs (tRNAs) at the wobble uridine 34 (U34) base are highly conserved and contribute to translation fidelity [36]. Partner enzymes in U34 tRNA modification, including ELP3 and CTU1/2 were found to be up-regulated in human breast cancers and sustain metastasis, through the translation of the oncoprotein DEK, that promotes the translation of the pro-invasive transcription factor LEF1 [37].

UCN (urocortin) has been previously reported to attenuate TGFβ1-induced *Snail1* and *Slug* expressions, both in ER+ BRCA and TNBC in vitro models [38], suggesting that urocortin may inhibit TGFβ1 oncogenic signaling and ultimately EMT. Therefore, *UCN* up-regulation would be associated with a better prognosis, a less invasive disease. However, it was also suggested that urocortin may have a dual role in cancer, since it differentially binds to CRFR1 or CRFR2 and either activates or blocks the Bcl-2/Bax/caspase-9 axis, leading to apoptosis or survival, respectively [39]. In clinical samples, *UCN* was found to be elevated in PRAD, although no correlations with clinical features were presented in this study [40].

The above described *SCARA5*, and also *SNCG*, *PARK2*, *CD300LG* and *KLHL4*, were down-regulated in both AR+ TNBC and PRAD.

The loss of epigenetic control of *SNCG* (synuclein-gamma) seems to be a molecular indicator of metastasis

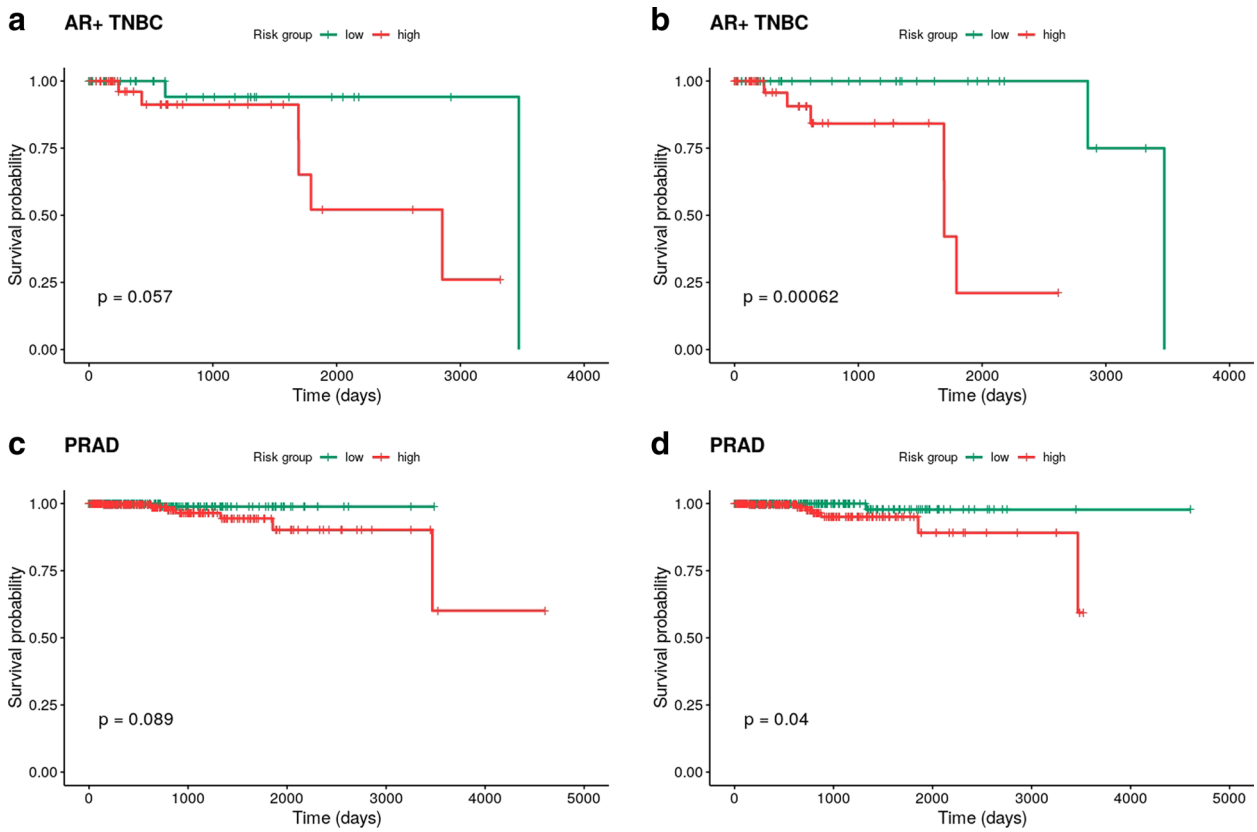

Fig. 8 Kaplan-Meier survival curves obtained for the AR+ TNBC and PRAD datasets using the variables selected by elastic net (EN) (**a** and **c**) and `twiner` (**b** and **d**). AR+ TNBC, androgen receptor positive Triple-Negative Breast Cancer; PRAD, Prostate Adenocarcinoma

in a wide range of human cancers, including BRCA and PRAD [41]. In this case, is the reactivation of *SNCG* gene expression by DNA demethylation the contributing factor to malignant progression of many solid tumors and its expression in primary carcinomas is an effective molecular indicator of distant metastasis. Silencing *SNCG* in a prostate cancer cell line has shown to decrease proliferation and invasion in vitro, and tumor growth in vivo, with the exception of castrated mice [42], suggesting AR-dependence. It was shown that *SNCG* interacts with AR and promotes prostate cancer cellular growth and proliferation by activating AR transcription in an androgen-dependent manner, whereas *SNCG* was almost undetectable in benign or androgen-independent tissues prostate lesions. *SNCG* in PRAD was also described to be activated by Cav-1 in the tumor microenvironment [43]. Nevertheless, decreased expression of *SNCG* may correspond to a more indolent disease. In breast cancer, it was shown that TNBC cell lines do not express *SNCG*, in accordance with our results [44]; although in another study using a small cohort of 55 cases there was no association between the clinicopathologic parameters including histologic grade, ER positivity and HER2 status and the level of *SNCG* [45].

PARK2 (PARKIN, E3 ubiquitin ligase) is involved in autosomal recessive parkinsonism. *PARK2* presents a partial mitochondrial localization at the outer mitochondrial membrane and its depletion results in abnormal mitochondrial morphology. An *in silico* analysis has shown that *PARK2* may be related to cell cycle control, suggesting a role in carcinogenic processes [46].

In accordance to our findings, *CD300LG* was found to be down-regulated in AR+ TNBC tissues when compared with adjacent normal studies [47], but its role in cancer is still unknown. This gene encodes the CD300 antigen-like family member G protein, also called nepmucin or CLM-9, expressed extensively in a variety of organisational venules and capillary endothelial cells in many organs. It is hypothesized that it may be involved in recruitment of immune cells, and that *CD300LG* down-regulation results in immune escape of cancer cells [48].

Also *KLHL4* (Kelch like family member 4) was expected to be down-regulated in breast cancer. A previous study has shown that this gene is down-regulated downstream of *IGFBP5*, silenced in response to stromal cells in ERα-positive breast cancer cells [49]. As this is part of a

mechanism of induced resistance to anti-estrogen therapy, *KLHL4* down-regulation is expected to be associated with worse prognosis.

From the biological interpretation above, several links between the results obtained by our method and disease biology have been established, which reinforces the ability of our method to identify shared disease features in breast and prostate cancers. Moreover, these genes are able to stratify patients into high or low risk, according to overall survival, and deserve further studies to clearly determine their role in the progression of BRAC and PRAD.

Conclusions

High-dimensional data leads to ill-posed inverse problems that cannot be tackled easily. Regularized optimization is a promising strategy to cope with undetermined problems, since it adds extra constraints to the loss function, which, if chosen carefully, can provide biological and clinical insight. We presented the `twiner` penalty, a correlation-based regularizer designed to enable the selection of similarly correlated genes in two diseases by sparse logistic regression, as a strategy to identify common key players in both diseases. The usefulness of the strategy proposed is shown in the context of Breast Invasive Carcinoma (BRCA) and Prostate Adenocarcinoma (PRAD), which show remarkable similarities at the hormone-related signaling pathways level. While being largely supported by the literature and clinical evidence by survival analysis, our results identified putative disease biomarkers which are expected to greatly improve our knowledge on the diseases and contribute to the definition of new target therapies.

Endnotes

[1] https://github.com/averissimo/brca.data/releases/download/1.0/brca.data_1.0.tar.gz

[2] https://github.com/averissimo/tcga.data/releases/download/2017.07.21-prad/prad.data_1.0.tar.gz

Abbreviations

AR: Androgen receptor; AUC: Area under the curve; BRCA: Breast invasive carcinoma; CDS: Coding sequence; EN: Elastic net; ER: Estrogen receptor; FPKM: Fragments per kilo base per million; HER2: Human epidermal growth factor receptor 2; MSE: Mean squared error; PC: Principal component; PCA: Principal component analysis; PR: Progesterone receptor; PRAD: Prostate adenocarcinoma; RNA-Seq: RNA sequencing; TCGA: The cancer genome atlas; TNBC: Triple-negative breast cancer; TWINER: Twin networks recovery

Acknowledgements

The authors thank André Veríssimo and Eunice Carrasquinha for helpful discussions during problem definition and data analysis.

Authors' contributions

MBL and SV designed the study, MBL implemented and performed the testings, MBL, SC, and SV analysed the results and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UID/EEA/50008/2019 (Instituto de Telecomunicações), UID/CEC/50021/2019 (INESC-ID), UID/EMS/50022/2019 (IDMEC, LAETA), PREDICT (PTDC/CCI-CIF/29877/2017), and PERSEIDS (PTDC/EMS-SIS/0642/2014). The funders had no role in the design of the study, collection, analysis and interpretation of data, or writing the manuscript.

Availability of data and materials

All the implementations described can be found in a R Markdown document available at https://github.com/sysbiomed/twiner, which allows full reproducibility and adaptation to new datasets.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

SV is member of the Editorial Board of BMC Bioinformatics. MBL and SC declare that they have no competing interests.

Author details

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal. [2]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol 9, 1000-029 Lisboa, Portugal. [3]Luis Costa Lab, Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Avenida Professor Egas Moniz, 1649-028 Lisboa, Portugal. [4]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal.

Received: 20 March 2019 Accepted: 6 June 2019
Published online: 25 June 2019

References

1. Risbridger GP, Davis ID, Birrell SN, Tilley WD. Breast and prostate cancer: more similar than different. Nat Rev Cancer. 2010;10:205–1.
2. van Hellemond IEG, Geurts SME, Tjan-Heijnen VCG. Current status of extended adjuvant endocrine therapy in early stage breast cancer. Curr Treat Options Oncol. 2018;19(5):26.
3. Arora K, Barbieri CE. Molecular subtypes of prostate cancer. Curr Oncol Rep. 2018;20:58.
4. Culig Z, Santer FR. Androgen receptor signaling in prostate cancer. Cancer Metastasis Rev. 2014;33:413.
5. Renoir J-M, Marsaud V, Lazennec G. Estrogen receptor signaling as a target for novel breast cancer therapeutics author links open overlay panel. Biochem Pharmacol. 2013;85(4):449–65.
6. Zazzo E, Galasso G, Giovannelli P, Donato M, Castoria G. Estrogens and their receptors in prostate cancer: Therapeutic implications. Front Oncol. 2018;8:2.
7. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med. 2010;363:1938–48.
8. Liu Y-X, Zhang K-J, Tang L-L. Clinical significance of androgen receptor expression in triple negative breast cancer - an immunohistochemistry study. Oncol Lett. 2018;15(6):10008–16.
9. Mina A, Yoder R, Sharma P. Targeting the androgen receptor in triple-negative breast cancer: current perspectives. OncoTargets Therapy. 2018;10:4675–85.
10. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1986;58(1):267–88.
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67(2):301–20.
12. Zhang W, Ota T, Chien VSJ, Wu B, Kuang R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol. 2013;9(3):1002975.
13. Veríssimo A, Oliveira AL, Sagot M-F, Vinga S. Degreecox - a network-based regularization method for survival analysis. BMC Bioinformatics. 2016;17(Supp 16):449.
14. Zhang W, Wan Yw, Allen GI, Pang K, Anderson ML, Liu Z. Molecular pathway identification using biological network-regularized logistic models. BMC Genomics. 2013;14(Suppl 8):7.

15. Alcaraz N, List M, Batra R, Vandin F, Ditzel HJ, Baumbach J. De novo pathway-based biomarker identification. Nucleic Acids Res. 2017;45(16): 151.
16. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, Raphael BJ, Marks DS, Ouellette BFF, Valencia A, Bader GD, Boutros PC, Stuart JM, Linding R, Lopez-Bigas N, Stein LD. Pathway and network analysis of cancer genomes. Nat Methods. 2015;12(7):615–21.
17. Lopes MB, Veríssimo A, Carrasquinha E, Casimiro S, Beerenwinkel N, Vinga S. Ensemble outlier detection and gene selection in triple-negative breast cancer data. BMC Bioinformatics. 2018;19:168.
18. Segaert P, Lopes MB, Casimiro S, Vinga S, Rousseeuw P. Robust identification of target genes and outliers in triple-negative breast cancer data. Stat Methods Med Res. 2018;0(0):1–15.
19. The Ensembl genome browser. http://www.ensembl.org/index.html. Accessed May 2017.
20. The Consensus CDS (CCDS) project. https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi. Release 20, Accessed May 2017.
21. Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted data. PLoS ONE. 2014;9(3):92209.
22. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in r. Bioinformatics. 2015;31(15):2595–7.
23. Cox DR. Regression models and life-tables. J R Stat Soc Ser B (Methodol). 1972;34(2):187–220.
24. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
26. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2017. R Foundation for Statistical Computing. https://www.R-project.org/.
27. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;7(43):47.
28. Gupta PB, Onder TT, Jiang G, Tao K, Kuperwasser C, Weinberg RA, Lander ES. Identification of selective inhibitors of cancer stem cells by high-throughput screening. Cell. 2009;138(4):645–59.
29. Yue R, Shen B, Morrison SJ. Clec11a/osteolectin is an osteogenic growth factor that promotes the maintenance of the adult skeleton. eLife. 2016;5: 18782.
30. Fan S, Liang Z, Gao Z, Pan Z, Han S, Liu X, Zhao C, Yang W, Pan Z, Feng W. Identification of the key genes and pathways in prostate cancer. Oncol Lett. 2018;16:6663–9.
31. Thompson JJ, Kaur R, Sosa CP, Lee J-H, Kashiwagi K, Zhou D, Robertson KD. ZBTB24 is a transcriptional regulator that coordinates with DNMT3B to control DNA methylation. Nucleic Acids Res. 2018;46(19):10034–51.
32. You K, Su F, Liu L, Lv X, Zhang J, Zhang Y, Liu B. SCARA5 plays a critical role in the progression and metastasis of breast cancer by inactivating the ERK1/2, STAT3, and AKT signaling pathways. Mol Cell Biochem. 2017;435:47–58.
33. Ulkera D, Ersoy YE, Gucin Z, Muslumanoglu M, Buyru N. Downregulation of SCARA5 may contribute to breast cancer via promoter hypermethylation. Gene. 2018;673:102–6.
34. Rahim B, O'Regan R. AR signaling in breast cancer. Cancers. 2017;9(3):21.
35. Lehmann BD, andXi Chen BJ, Estrada MV, Johnson KN, Shyr Y, Moses HL, Sanders ME, Pietenpol JA. Refinement of triple-negative breast cancer molecular subtypes: Implications for neoadjuvant chemotherapy selection. PLoS ONE. 2016;11(6):0157368.
36. Dewez M, Bauer F, Dieu M, Raes M, Vandenhaute J, Hermand D. The conserved Wobble uridine tRNA thiolase Ctu1-Ctu2 is required to maintain genome integrity. PNAS. 2008;105(14):5459–64.
37. Delaunay S, Rapino F, Tharun L, Zhou Z, Heukamp L, Termathe M, Shostak K, Klevernic I, Florin A, Desmecht H, Desmet CJ, Nguyen L, Leidel SA, Willis AE, Büttner R, Chariot A, Close P. Elp3 links tRNA modification to IRES-dependent translation of LEF1 to sustain metastasis in breast cancer. J Exp Med. 2016;213(11):2503–23.
38. Jin L, Zhu C, Wang X, Li C, Cao C, Yuan J, Li S. Urocortin attenuates TGFβ1-induced snail1 and slug expressions: Inhibitory role of smad7 in smad2/3 signaling in breast cancer cells. PLoS ONE. 2015;16(11):2494–503.
39. Jin L, Zhang Q, Guo R, Wang L, Wang J, Wan R, Zhang R, Xu Y, Li S. Different effects of corticotropin-releasing factor and urocortin 2 on apoptosis of prostate cancer cells in vitro. J Mol Endocrinol. 2011;47(2): 219–27.
40. Arcuri F, Cintorino M, Florio P, Floccari F, Pergola L, Romagnoli R, Petraglia F, Tosi P, Vecchio MTD. Expression of urocortin mRNA and peptide in the human prostate and in prostatic adenocarcinoma. Prostate. 2002;52(3):167–72.
41. Liu H, Liu W, Wu Y, Zhou Y, Xue R, Luo C, Wang L, Zhao W, Jiang JD, Liu J. Loss of epigenetic control of synuclein-gamma gene as a molecular indicator of metastasis in a wide range of human cancers. Cancer Res. 2005;65(17):7635–43.
42. Chen J, Jiao L, Xu C, Yu Y, Zhang Z, Chang Z, Deng Z, Sun Y. Neural protein gamma-synuclein interacting with androgen receptor promotes human prostate cancer progression. BMC Cancer. 2012;12:593.
43. Ayala G, Morello M, Frolov A, You S, Li R, Rosati F, Bartolucci G, Danza G, Adam RM, Thompson TC, Lisanti MP, Freeman MR, Vizio DD. Loss of caveolin-1 in prostate cancer stroma correlates with reduced relapse-free survival and is functionally relevant to tumour progression. J Pathol. 2013;231(1):77–87.
44. Tian L, Zhao Y, Truong M-J, Lagadec C, Bourette RP. Synuclein gamma expression enhances radiation resistance of breast cancer cells. Oncotarget. 2018;9(44):27435–47.
45. Cirak Y, Furuncuoglu Y, Yapicier O, Alici S, Argon A. Predictive and prognostic values of bubR1 and synuclein-gamma expression in breast cancer. Int J Clin Exp Pathol. 2015;8(5):5345–53.
46. Salazar C, Ruiz-Hincapie P, Ruiz LM. The interplay among PINK1/PARKIN/Dj-1 network during mitochondrial quality control in cancer biology: Protein interaction analysis. Cells. 2018;7(154).
47. Shen X, Xie B, Ma Z, Yu W, Wang W, Xu D, Yan X, Chen B, Yu L, Li J, Chen X, Ding K, Cao F. Identification of novel long non-coding RNAs in triple-negative breast cancer. Oncotarget. 2015;6(25):21730–9.
48. Wang Q, Liu Y, Chen Y, Wang K, Xie W, Wei D, Hu L. CD300LG improves the cytotoxic activity of CIK. Cent Eur J Immunol. 2017;42(2):117–122.
49. Leyh B, Dittmer A, Lange T, Martens JW, Dittmer J. Stromal cells promote anti-estrogen resistance of breast cancer cells through an insulin-like growth factor binding protein 5 (IGFBP5)/B-cell leukemia/lymphoma 3 (Bcl-3) axis. Oncotarget. 2015;6(36):39307–28.

Publisher's Note