BMC Medical Genomics

RESEARCH

Open Access

Highly robust model of transcription regulator activity predicts breast cancer overall survival

Chuanpeng Dong[1,2†], Jiannan Liu[2†], Steven X. Chen[1], Tianhan Dong[3], Guanglong Jiang[1,2], Yue Wang[1], Huanmei Wu[2], Jill L. Reiter[1*] and Yunlong Liu[1,2*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2019
Columbus, OH, USA. 9-11 June 2019

Abstract

Background: While several multigene signatures are available for predicting breast cancer prognosis, particularly in early stage disease, effective molecular indicators are needed, especially for triple-negative carcinomas, to improve treatments and predict diagnostic outcomes. The objective of this study was to identify transcriptional regulatory networks to better understand mechanisms giving rise to breast cancer development and to incorporate this information into a model for predicting clinical outcomes.

Methods: Gene expression profiles from 1097 breast cancer patients were retrieved from The Cancer Genome Atlas (TCGA). Breast cancer-specific transcription regulatory information was identified by considering the binding site information from ENCODE and the top co-expressed targets in TCGA using a nonlinear approach. We then used this information to predict breast cancer patient survival outcome.

Result: We built a multiple regulator-based prediction model for breast cancer. This model was validated in more than 5000 breast cancer patients from the Gene Expression Omnibus (GEO) databases. We demonstrated our regulator model was significantly associated with clinical stage and that cell cycle and DNA replication related pathways were significantly enriched in high regulator risk patients.

Conclusion: Our findings demonstrate that transcriptional regulator activities can predict patient survival. This finding provides additional biological insights into the mechanisms of breast cancer progression.

Keywords: Breast cancer, Transcription regulators, Prognostic model

Background

Breast cancer is the most frequently diagnosed cancer and the second leading cause of cancer deaths in women worldwide [1]. In the United States, breast cancer accounted for 30% of all new cancer cases and 14% of cancer deaths for women in 2017, with an estimated 252,710 newly diagnosed cases and 40,610 deaths [2]. However, because of intertumoral and intratumor heterogeneity, significant challenges exist in designing effective treatment regimens and in predicting clinical outcomes.

Currently, breast tumor classification is primarily based on histopathologic features and the expression of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) [3]. These subtypes differ with respect to available receptor-targeted therapies, response to treatment, clinical outcomes and risk of acquiring resistance to therapy [4]. Hundreds of other biomarkers have been reported in breast cancer for prognostic

* Correspondence: jireiter@iu.edu; yunliu@iu.edu
†Chuanpeng Dong and Jiannan Liu contributed equally to this work.
[1]Department of Medical and Molecular Genetics, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
Full list of author information is available at the end of the article

and therapeutic applications [5]. Over the past decade, multi-gene signatures have been developed for breast cancer subtyping and risk stratification [6]. For instance, the PAM50 gene signature measures the expression levels of 50 genes in breast cancer samples to classify a tumor as one of five intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like and normal like), and it has prognostic value in both untreated and tamoxifen treated patient populations [7, 8]. The MammaPrint assay categorized patients into good or poor risk groups using 70 genes and has been approved by the Food and Drug Administration (FDA) to aid in predicting prognosis for breast cancer patients with specific clinical characteristics [9, 10]. However, these gene-based risk models have certain limitations and to date, there is no multigene test that has been approved for recommending adjuvant treatment for triple-negative (ER/PR/HER2-negative) breast tumors. There remains a critical need for the development of a robust model that can aid in effectively predicting individual patient prognosis for hormone-receptor-negative breast tumors that can convey additional biological information from gene expression profiles.

Transcriptional regulators, including the transcription factors, cofactors, chromatin remodelers, histone modification proteins and other DNA binding proteins, play fundamental roles in many cellular processes including response to extracellular and intracellular signals. By binding to promoter regions of genes, transcription regulators are able to up- or down-regulate specific target genes, thereby affecting many cellular activities. In recent years, several studies have implemented machine-learning methods in cancer prognostic biomarker development. For example, a support vector machine (SVM)-classifier has been used to construct a breast cancer prognostic signature consisting of 10 microRNAs that accurately predicted breast cancer stage [11]. Moreover, a clustering-based method identified microRNA combinatorial biomarkers with high accuracy and efficiency [12]. However, there are challenges with using machine-learning methods with high-dimensional profiles [13].

We hypothesized that transforming gene expression profiles into transcription regulator activity levels would reduce the dimensionality while retaining the most useful information from the gene expression profile during feature selection and the model training process. Herein, we built a transcription regulator-based model by mining gene expression profiles from 1097 breast cancer patients obtained from The Cancer Genome Atlas (TCGA). We implemented a rank-based score function to estimate the regulator activity levels to build the prediction model. In addition, the predictive power of this transcription regulator activity-based model was validated on over 5000 breast cancer patient profiles in the Gene Expression Omnibus (GEO) database. Gene set enrichment analysis demonstrated that the transcriptional regulator-based prognostic signature identified key pathways involved in breast cancer development and progression. Thus, transcription regulator activity models are expected to provide new information that could be used to develop new treatment strategies for breast cancer.

Methods

Gene expression profiles

Breast cancer gene expression datasets and the corresponding clinical data were downloaded from TCGA and GEO databases. TCGA breast cancer RNA-sequencing data (log2(norm_count + 1) normalized) [14], including the corresponding clinical information, were downloaded from UCSC Xena (http://xena.ucsc.edu, version 2017-10-13) [15]. Validation microarray datasets from 29 studies were downloaded from GEO (Additional file 1: Table S1). For the Affymetrix microarray studies, raw data of microarray datasets were downloaded in CEL format or processed soft matrix files were used for datasets without CEL. After background correction, the robust multi-array average (RMA) method was used to normalize the Affymetrix microarray data [16]. For microarray studies using other platforms, the series matrix files were downloaded from the NCBI GEO website [17]. Probes were annotated to the gene symbols, and multiple probes annotated to the same gene were merged and mean values were calculated as the expression of the corresponding genes. The clinical data were acquired from the respective literature citations.

Chip-seq datasets from ENCODE

The transcriptional regulators and their target genes list was retrieved from the ENCODE project via ChIPbase V2 (http://rna.sysu.edu.cn/chipbase/) [18, 19]. The transcription factor and chromatin remodeling factor regulatory domains were defined using the following settings: 1 kb upstream and downstream of the target gene transcription start sites, union combination mode and all motifs. The resulting lists were then limited to ENCODE as the source. In total, 180 transcription factors or other chromatin remodeling factor data were included and together, are referred to as transcription regulators in this study.

Survival analysis and statistical methods

The Cox proportional-hazards model was used to select the candidate transcription regulator that significantly associated with patients' overall survival [20, 21]. A multivariate Cox regression was performed

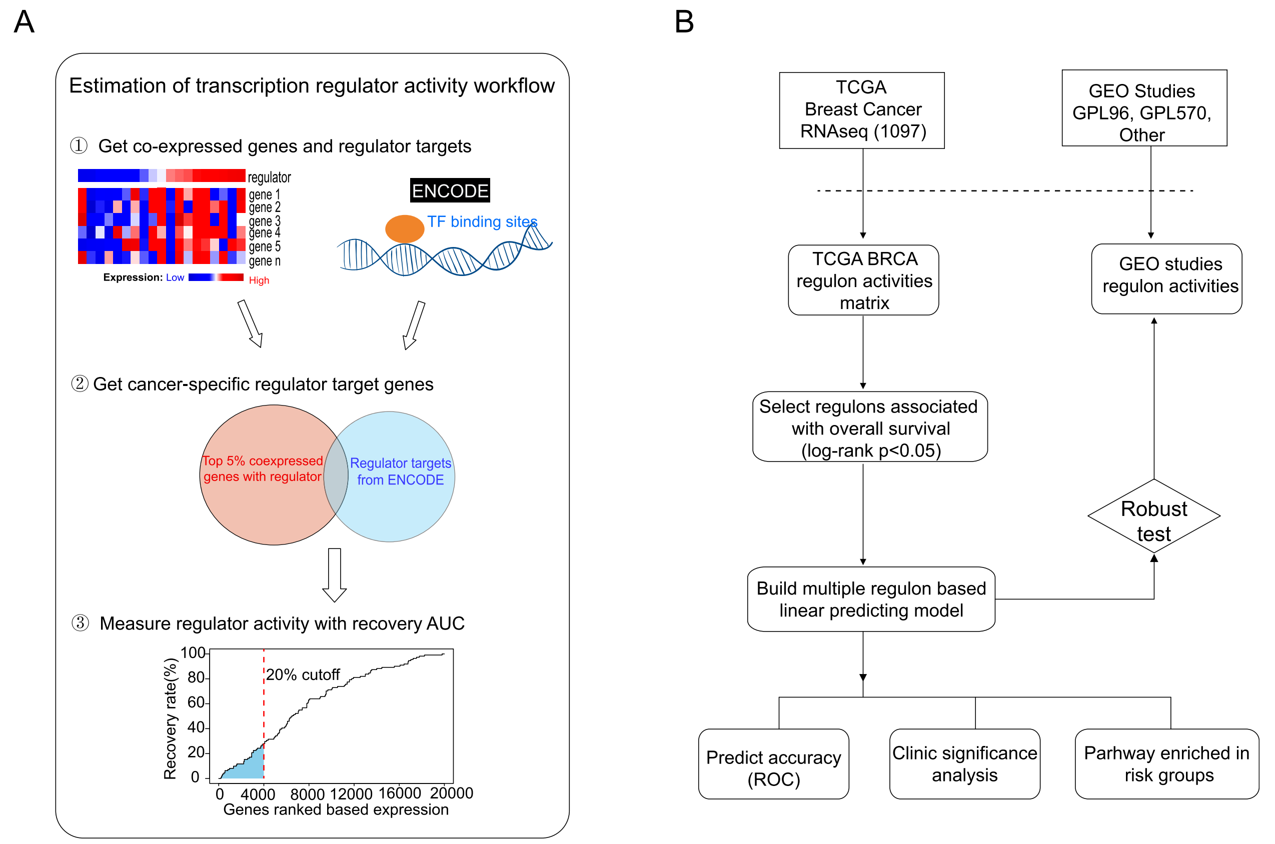
A
Estimation of transcription regulator activity workflow
① Get co-expressed genes and regulator targets
regulator
gene 1
gene 2
gene 3
gene 4
gene 5
gene n
Expression: Low High
ENCODE
TF binding sites
② Get cancer-specific regulator target genes
Top 5% coexpressed genes with regulator
Regulator targets from ENCODE
③ Measure regulator activity with recovery AUC
Recovery rate(%)
100
80
60
40
20
0
20% cutoff
0 4000 8000 12000 16000 20000
Genes ranked based expression
B
TCGA Breast Cancer RNAseq (1097)
GEO Studies GPL96, GPL570, Other
TCGA BRCA regulon activities matrix
GEO studies regulon activities
Select regulons associated with overall survival (log-rank p<0.05)
Robust test
Build multiple regulon based linear predicting model
Predict accuracy (ROC)
Clinic significance analysis
Parhway enriched in risk groups

Fig. 1 Workflow of the study. **a**. Estimation of transcription regulator activities workflow. Co-expression analysis of each regulator with all genes in TCGA breast cancer cohort was performed using GRNBoost2 and regulator binding site information was retrieved from ENCODE (step 1). The intersection of the top 5% of co-expressed genes with regulator target genes was identified (step 2) and used to calculate the regulator activity with a rank-based approach (step 3). **b**. Overall workflow of the study. Abbreviations: TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus

to weight each of the selected potential transcription regulators with an adapted coefficient. A transcriptional regulator activity based risk score formula was constructed by including statistically significant genes weighted by their estimated multivariable Cox's regression adapted coefficients [22].

$$\text{Risk score} = \sum\nolimits_1^n \mathit{coefficient}(i) * \mathit{transcription\ regulator}(i)\ \mathit{activity}.$$

Patients were then divided into high-risk or low-risk groups using the median risk score as the cutoff [23]. The Kaplan-Meier curve was used to compare the survival probabilities between the high-risk and low-risk groups, and the log-rank test was adopted to test the difference in survival rate between patients in the high- and low-risk subgroups. A *p*-value less than 0.05 was considered as statistically significant. The survival analyses were conducted with the *survival* package in the statistical environment R (v3.5.1). Statistical computing and visualization were conducted with R.

Gene set enrichment analysis was conducted with GSEA software (http://www.broadinstitute.org/gsea) [24] using the canonical pathways collection (version, c2.cp.v6.2) [25] and a false discovery rate (FDR) value less than 0.01 after performing 1000 permutations was considered to be significant.

Results

Dataset and workflow for estimating transcription regulator activities

RNA-seq profiles from 1097 primary breast cancer patients were used as the training set. Transcription regulator binding site information was extracted from ChIP-seq data contained within ENCODE. This data included 180 regulators and their gene targets. Additionally, 29 breast cancer studies in GEO were used as a validation set (Additional file 1: Table S1).

We estimated transcription regulator activity using a method that was adapted from the single cell algorithm SCENIC [26]. The workflow for obtaining this estimate followed three steps as shown in Fig. 1.

Firstly, co-expression analysis for each of the 180 transcription regulators with all genes in the TCGA breast cancer sets was conducted using GRNBoost2, an efficient algorithm for regulatory network inference using gradient boosting. Secondly, the intersection of the top 5% of transcription regulator-gene pairs (i.e, regulons) from GRNBoost2 [27] and genes from the ENCODE ChIP-seq dataset produced the set of genes considered to be the regulator targets that were expressed in breast cancers. Thirdly, the activity of each transcription regulator in each patient was measured with the rank based approach AUCell package [28]. We calculated the enrichment of the identified breast cancer specific regulator target as an area under the recovery curve across the ranking of all genes in a particular patient's profile. The output is the enrichment matrix of transcription regulator activity for each patient; genes were ranked by their expression value and the cutoff parameter was set as 0.2 for both the training and testing sets. The rank-based method enabled estimation of transcription regulator activity at the individual patient level.

Transcription regulator activities correlated with breast cancer subtypes

We postulated that transcription regulator activity might be similar in breast cancers with similar gene expression profiles. This would suggest that breast tumors within the same molecular subtypes might have similar transcription regulator profiles. To address this question, we asked whether a correlation existed between commonly used molecular PAM50 subtypes and the transcriptional regulator activity. Each transcriptional regulator activity was calculated by the approach described above. We then conducted an unsupervised clustering with the regulator activity matrix. These results indicated that transcription regulator activity specifically identified the basal-like breast tumors (Fig. 2). Transcriptional regulators that exhibited higher activities in the basal-like tumors compared to the other breast cancer subtypes include BRF1, CTCFL, E2F1, FOXM1, GTF2B, GTF3C2, HCFC1, KAT2A, MEF2C, MYBL2, MYC, POLR3G and WRNIP1. These findings are consistent with a previous report that FOXM1 functions as a specific marker for triple negative breast cancer [29]. In addition, previous studies have demonstrated that triple-negative tumors exhibit elevated expression of MYC regulatory genes and increased activity of the MYC pathway [30].

Transcriptional regulator activity was associated with breast cancer prognosis

We next asked whether transcription regulator activity would be useful in predicting breast cancer patient

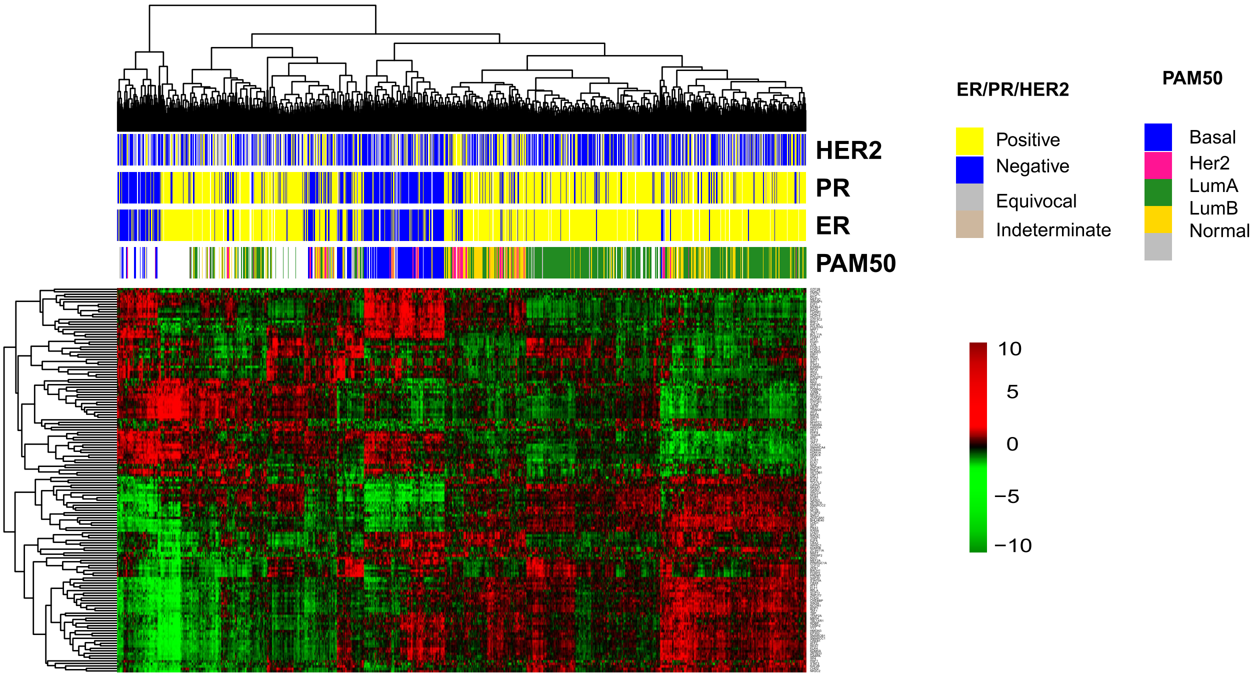

Fig. 2 Correlation of transcription regulator activities with breast cancer subtypes. The distribution of transcription regulator activity and breast cancer subtype with hormone receptor status and PAM50 molecular classification. The transcription regulator activity profile was generated by unsupervised cluster analysis, in which the rows represented the regulator activities and the columns represented the samples. The transcription regulator activity score varies between − 10 and + 10 as indicated by a gradient from green to red color

Table 1 Transcriptional regulator activity associated with breast cancer overall survival

Regulator	log-rank p	HR	95% CI	*p* value
BATF	3.00E-06	0.45	(0.32–0.64)	5.26E-06
ESR1	0.002	0.60	(0.43–0.83)	0.0023
IRF1	0.0034	0.62	(0.44–0.85)	0.0037
THAP1	0.0099	1.53	(1.10–2.12)	0.0105
KDM1A	0.0158	0.67	(0.48–0.93)	0.0165
TBP	0.0167	1.49	(1.07–2.06)	0.0173
JUNB	0.0199	0.68	(0.49–0.94)	0.0206
ATF3	0.021	0.69	(0.50–0.95)	0.0218
CHD7	0.0236	1.45	(1.05–2.01)	0.0243
STAT2	0.032	0.70	(0.51–0.97)	0.0329
GTF2B	0.04	0.71	(0.52–0.99)	0.0409
IKZF1	0.0406	0.71	(0.52–0.99)	0.0416
FOXA1	0.0459	0.72	(0.52–1.00)	0.0469
MAX	0.046	0.72	(0.52–1.00)	0.0469
REST	0.0493	1.38	(1.00–1.91)	0.0502

**HR* Hazard ratio, *CI* confidence interval

prognosis. To address this question, we applied the Cox proportional-hazards model to screen transcription regulators that correlated with the overall survival of breast cancer patients. We found that fifteen transcription regulator activities showed significant associations with overall survival (Table 1). Among those 15 regulator activities, four transcriptional regulators had hazard ratio above 1 (CHD7, REST, TBP and THAP1), indicating that elevated activities of these transcriptional regulators were associated with poor prognosis. Eleven transcription regulators (ATF3, BATF, ESR1, FOXA1, GTF2B, IKZF1, IRF1, JUNB, KDM1A, MAX and STAT2) had a hazard ratio less than 1, suggesting that higher activity of these transcription regulators in breast tumors might be beneficial for patient survival.

Transcriptional regulator model predicted breast cancer overall survival

We then calculated risk scores for each breast cancer patient in the training set as described in the Methods section. Patients were divided into low-risk and high-risk groups using median risk score as the cut-off. The risk scores distribution, survival status and transcription regulator profiles for the TCGA training set are shown

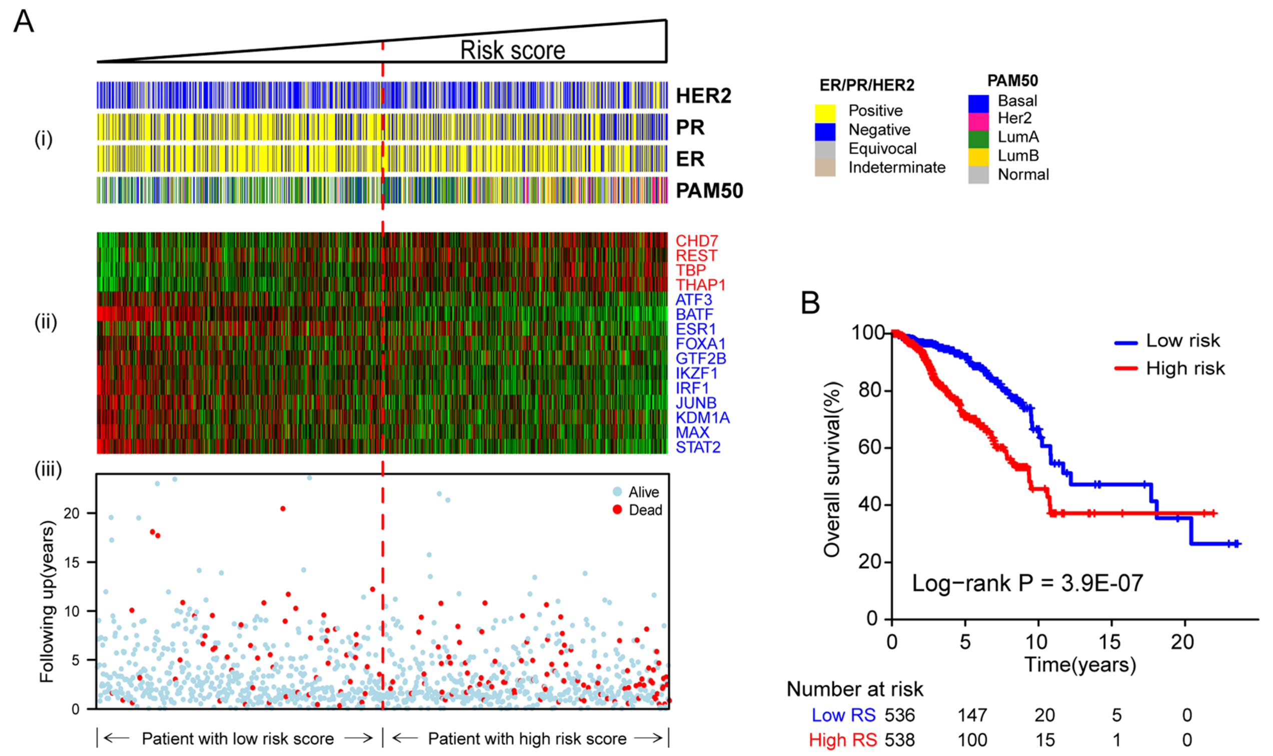

Fig. 3 Regulator based risk model of TCGA breast cancer patients. **a** The distribution of the significant regulator activities, patients' survival status and gene expression signature were analyzed in the TCGA breast cancer patients. (i) Hormone receptor status and PAM50 molecular classification of breast cancer patients. (ii) Heatmap of the selected regulator activities profile. (iii) Patients overall survival status and time. Rows represent genes, and columns represent patients. The red dashed line represents the risk score median cutoff dividing patients into low-risk and high-risk groups. **b** Kaplan-Meier estimates of patient survival in high- and low-risk groups based on transcriptional regulator activities

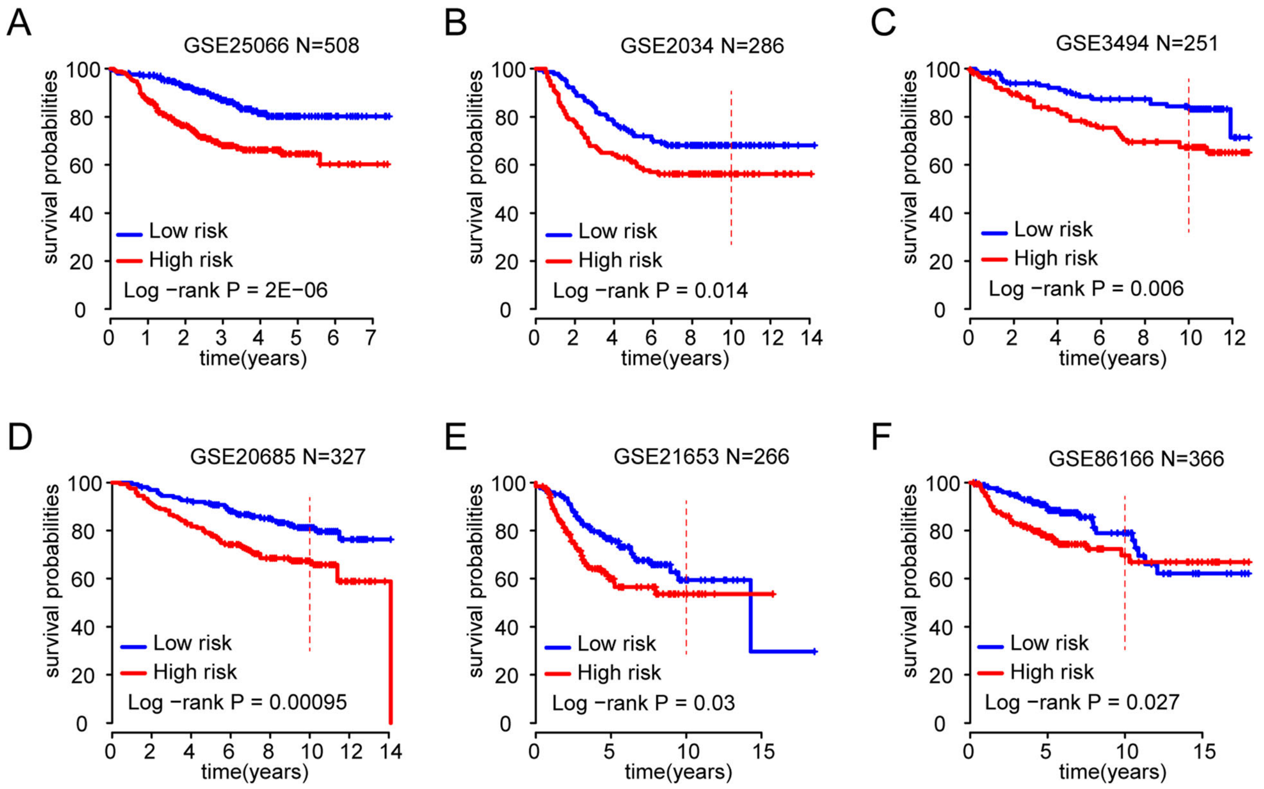

Fig. 4 Validation of the transcriptional regulator model with GEO datasets. Kaplan-Meier estimates of the survival of independent breast cancer datasets using the regulator activity signature. **a** Kaplan-Meier curves of overall survival for **a** GSE25066, **b** GSE2034, **c** GSE3494, **d** GSE20685, **e** GSE21653 and **f** GSE86166. The low-risk and high-risk groups of patients was determined on the basis of the median risk score for each validation dataset. The tick marks on the Kaplan-Meier curves represent censored subjects. The statistically significant differences between the two curves were determined by the two-sided log-rank test. The vertical red dashed line indicates the 10-year mark that was used for interpreting survival difference between groups

in Fig. 3a. Our results show that patients in the high-risk group had significantly shorter overall survival time than those in the low-risk group (log-rank $P < 0.0001$) (Fig. 3b).

In addition, we found that more hormone receptor negative (ER- and PR-negative) breast cancer patients were assigned to the high-risk group. Likewise, the basal-like and HER2-enriched breast cancer patients also were in the high-risk group, which is consistent with the clinical findings [31] that HER2-enriched and lumB subtypes of breast cancer have a poor prognosis compared with the lumA and normal-like breast cancers.

Validation of the transcription regulator model with independent datasets

To test whether the high performance of the risk-score model in the training dataset might have resulted from overfitting, we evaluated the performance of the transcription regulator activity model using independent breast cancer datasets from the GEO. Consistent with the training set, patients with high risk scores showed a clear trend of decreased survival rate in 23 of 29 GEO studies (Fig. 4 a-f and Additional file 1: Figure S1).

Notably, these independent validation GEO sets used different gene-expression platforms, such as the Affymetrix Human Genome U133A/B Array (Fig. 4 a-c), Affymetrix Human Genome U133 Plus 2.0 Array (Fig. 4 d and e) and the Illumina platform (Fig. 4f). These data strongly suggest that patients assigned to the high-risk group had significantly decreased 10-year survival, compared with those patients in the low risk groups. Taken together, we conclude that this risk-score model shows robust performance across datasets and platforms. We also tried to compare our regulator model with an existing genomic classification assay. MammaPrint risk for each patient in the TCGA training set and the GEO validation breast cancer set was measured with the genefu package in R 3.6.0 [32]. The results are shown in Additional file 1: Figure S3. The ROC curve showed that our regulator-based model performed better than the MammaPrint

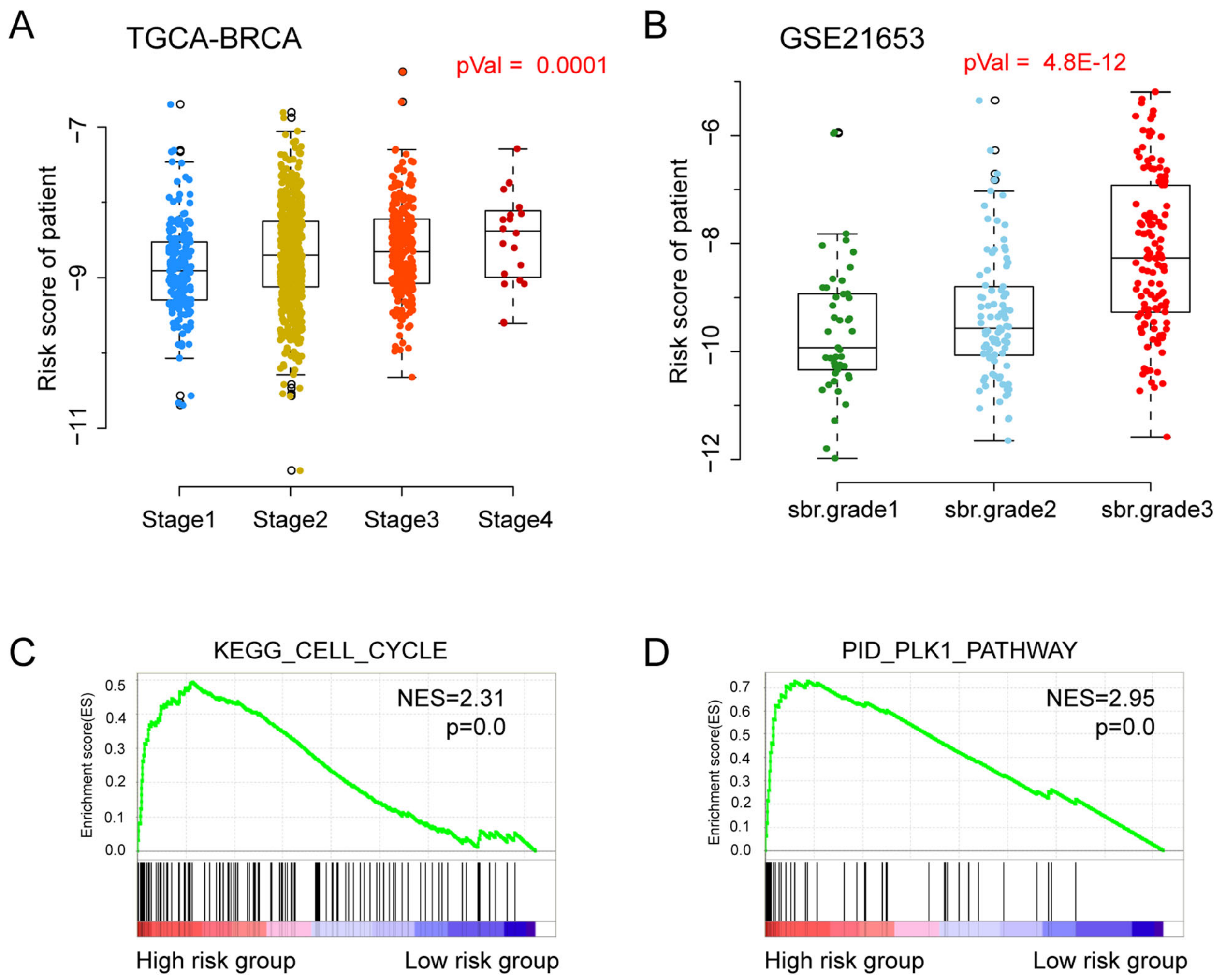

Fig. 5 Association of the transcription regulator-based risk score with pathological and molecular features. The regulator risk score is shown for breast cancer stages in the TCGA cohort (**a**) and in a selected validation set GSE21653 (**b**). One-way ANOVA *p* values are provided. Cancer related pathways that were significantly altered in patients with high regulator risk scores included cell cycle (**c**) and PLK1 signaling pathways (**d**). Normalized enrichment score (NES) was used to evaluate the enrichment results

risk model in the two largest testing sets GSE25066 (Additional file 1: Figure S3A) and GSE20685 (Additional file 1: Figure S3B).

Association of transcription regulator model with clinical and molecular features

To explore potential molecular mechanisms that might contribute to the clinical association with the transcription regulator activity-based model, we analyzed the correlation between the regulator risk score and the cancer clinical stage. We found the transcription regulator-based risk score and the pathological stage of breast cancer patients was correlated in both the training and validation sets (Fig. 5 a and b), where the one-way ANOVA *p* value reached 0.0001 for TCGA and p value was 4.8E-12 for GSE21653.

In addition, we performed Gene Set Enrichment Analysis (GSEA) on the TCGA breast cancer cohort. Compared with the patients assigned as low-risk, the high-risk patient group showed that the top two enriched pathways were cell cycle (Fig. 5c) and PLK signaling (Fig. 5d), which both play critical roles in cancer initiation and development. These findings provide evidence that the risk score can provide information relevant to potential molecular mechanisms that might be involved in tumor progression and survival outcome in breast cancer patients.

Discussion

In this study, a transcriptional regulator-based model was established for predicting the prognosis of breast cancer. By transforming gene expression profiles to transcription regulator activity levels, we demonstrated that regulon activity can be used to explore breast cancer gene expression data. After identifying that regulons significantly associated with breast cancer overall survival, we further constructed a multi-regulator activity-based prediction model for breast cancer. The finding from the training set was validated in different independent GEO sets (> 3500 patients), which indicates that the regulator activity-based model is robust.

Among the regulators that were used in the model, several are known to play a role in breast cancer. The chromatin remodeler CHD7 is one of the most commonly amplified CHD genes in breast cancer, and mRNA expression levels of CHD7 are significantly up-regulated in basal-like breast cancer [33]. The histone demethylase KDM1A, which functions to repress and activate transcription by mediating histone H3K4me1/2 and H3K9me1/2 demethylation, respectively, was reported to be present at significantly lower levels in breast cancer samples compared with normal tissues [34, 35]. Consistent with the previous literature, we found KDM1A was significantly lower in those patients with worse survival. Silencing of THAP–zinc finger protein THAP1 inhibits endothelial G1/S cell-cycle progression [36]. High FOXA1 was associated with better breast cancer specific survival among ER-positive breast cancer [37]. Functional studies have demonstrated that JUNB plays a pro-survival role in breast cancer cells in response to a lethal dose of flavopiridol [38]. Low mRNA expression of ESR1 is a determinant of tamoxifen resistance in ER-positive breast cancer [39]. The majority of the selected transcription regulators had been demonstrated previously to be strongly associated with breast cancer progression; therefore, differences in crucial regulator activity between low- and high-risk groups may provide new information for better understanding breast tumor pathology and risk for recurrence.

We hypothesized that limiting a prognostic gene signature to transcription regulators would reduce the dimensionality of the expression profile during feature selection and model training process. Additionally, we predicted that such a transcription regulator activity profile would be less susceptible to variability in the expression of individual genes and may thereby improve the prognostic significance of tumor gene profiling. To the best of our knowledge, a breast cancer prognostic signature with high prediction power has not yet been constructed using only information provided by transcription regulatory factors. Compared with other published molecular signatures and panels for breast cancer, this transcriptional regulator-based signature was highly robust across different datasets and platforms with very large-scale breast cancer samples, as the TCGA data was from RNA-sequencing, while various microarray platforms were used in GEO. The robustness of this model arises primarily from two aspects. Firstly, each transcriptional regulator activity was estimated using hundreds of direct target genes by considering both co-expression and ENCODE results, which produced a highly stable model without perturbation due to the variability of single gene expression levels. Secondly, the transcription regulator activity estimation was based on the rank of the absolute expression value of its target genes, which enabled it to perform well across different gene expression platforms. In addition, the transcriptional regulator activities were estimated in a tissue-specific manner, which allows the regulator activity based prognostic risk score to deliver additional biological information from breast cancer mRNA expression profiles.

Limitations of this study include the assumption that mRNA expression levels from the different platforms were measured appropriately and reflect the actual mRNA abundance of each gene. Secondly, the binding targets of the regulators obtained from ENCODE were not specific to breast cancer, so it is possible that a given regulator might not bind to all of the same targets in breast tumors. Thirdly, the cutoff for each dataset was determined separately based on the median risk score of the respective dataset; however, we did find that the value of the risk scores were in a similar range (– 12 to – 5).

Conclusion

In the present study, we built a robust model for predicting overall survival based on biologically relevant transcriptional regulator information and further validated it using large cohorts of breast cancer patients. The transcription regulator model should enhance our understanding of breast cancer progression and guide personalized treatment selection.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-020-0688-z.

Additional file 1: Table S1. GEO datasets used in this study. **Figure S1.** Transcriptional regulator model predicts survival in breast cancer datasets. **Figure S2.** The regulators risk groups associated with average survival time. **Figure S3.** Comparison of the performance of the regulator model with MammaPrint.

Acknowledgements
Not applicable

About this supplement
This article has been published as part of *BMC Medical Genomics Volume 13 Supplement 5, 2020: The International Conference on Intelligent Biology and Medicine (ICIBM) 2019: Computational methods and application in medical genomics (part 1)*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-13-supplement-5.

Authors' contributions

CD and JL participated in the study design, performed the majority of the experiments, analyzed the data and drafted the manuscript. SXC, TD, GJ, YW and HW provided intellectual discussions and suggestions. JLR participated in writing the manuscript. YL designed and directed the project. All authors read and approved the final manuscript.

Funding

This study is in part supported by the Indiana University Grand Challenge Precision Health Initiative.

Publication costs are funded by Indiana University Grand Challenge Precision Health Initiative.

Availability of data and materials

The datasets analyzed in the current study are publicly available in TCGA (https://portal.gdc.cancer.gov and https://tcga.xenahubs.net/download/TCGA.BRCA.sampleMap/HiSeqV2.gz). The microarray studies analyzed during the current study, GSE25066(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse25066), GSE2034(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034), GSE3494(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse3494), GSE20685(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse20685), GSE21653(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse21653) and GSE86166(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse86166), are available in the GEO Datasets repository.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Medical and Molecular Genetics, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [2]Department of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA. [3]Department of Pharmacology and Toxicology, Indiana University School of Medicine, Indianapolis, IN 46202, USA.

Published: 3 April 2020

References

1. Tao Z, Shi A, Lu C, Song T, Zhang Z, Zhao J. Breast cancer: epidemiology and etiology. Cell Biochem Biophys. 2015;72(2):333–8.
2. DeSantis CE, Ma J, Sauer AG, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin. 2017;67(6):439–48.
3. van de Ven S, Smit VT, Dekker TJ, Nortier JW, Kroep JR. Discordances in ER, PR and HER2 receptors after neoadjuvant chemotherapy in breast cancer. Cancer Treat Rev. 2011;37(6):422–30.
4. Polyak K. Heterogeneity in breast cancer. J Clin Invest. 2011;121(10):3786–8.
5. Lee E, Moon A. Identification of biomarkers for breast cancer using databases. J Cancer Prev. 2016;21(4):235–42.
6. Győrffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res. 2015;17(1):11.
7. Martín M, Prat A, Rodríguez-Lescure Á, Caballero R, Ebbert MT, Munárriz B, et al. PAM50 proliferation score as a predictor of weekly paclitaxel benefit in breast cancer. Breast Cancer Res Treat. 2013;138(2):457–66.
8. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna breast cancer prognostic gene signature assay and nCounter analysis system using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer. 2014;14(1):177.
9. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415:530–6.
10. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med. 2009;360(8):790–800.
11. Sathipati SY, Ho SY. Identifying a miRNA signature for predicting the stage of breast cancer. Sci Rep. 2018;8(1):16138.
12. Yang Y, Huang N, Hao L, Kong W. A clustering-based approach for efficient identification of microRNA combinatorial biomarkers. BMC Genomics. 2017; 18(2):210.
13. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. 2017;7(1):11707.
14. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
15. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, et al. The UCSC cancer genomics browser: update 2015. Nucleic Acids Res. 2014; 43(D1):D812–7.
16. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
17. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41(D1):D991–5.
18. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306(5696):636–40.
19. Zhou KR, Liu S, Sun WJ, Zheng LL, Zhou H, Yang JH, et al. ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. Nucleic Acids Res. 2017;45(D1): D43–50. https://doi.org/10.1093/nar/gkw965
20. Cox DR. Regression models and life-tables. J R Statist Soc B. 1972;34(2):187–220.
21. Sahar M.A. Mahmoud, Emma Claire Paish, Desmond G. Powe, R. Douglas Macmillan, Matthew J. Grainge, Andrew H.S. Lee, Ian O. Ellis, Andrew R. Green, Tumor-Infiltrating CD8 Lymphocytes Predict Clinical Outcome in Breast Cancer. Journal of Clinical Oncology 2011;29(15):1949–55.
22. Valk PJ, Verhaak RG, Beijen MA, Erpelinck CA, Barjesteh van Waalwijk van Doorn-Khosrovani S, Boer JM, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med. 2004;350(16):1617–28.
23. Zheng S, Zheng D, Dong C, Jiang J, Xie J, Sun Y, et al. Development of a novel prognostic signature of long non-coding RNAs in lung adenocarcinoma. J Cancer Res Clin Oncol. 2017;143(9):1649–57.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
25. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
26. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083–6.
27. Ghahramani A, Watt FM, Luscombe NM. Generative adversarial networks simulate gene expression and predict perturbations in single cells. BioRxiv. 2018:262501.
28. Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M, et al. Toward minimal residual disease-directed therapy in melanoma. Cell. 2018; 174(4):843–55 e819.
29. Tan Y, Wang Q, Xie Y, Qiao X, Zhang S, Wang Y, et al. Identification of FOXM1 as a specific marker for triple-negative breast cancer. Int J Oncol. 2019;54(1):87–97.
30. Horiuchi D, Kusdra L, Huskey NE, Chandriani S, Lenburg ME, Gonzalez-Angulo AM, et al. MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. J Exp Med. 2012;209(4): 679–96.
31. Liu MC, Pitcher BN, Mardis ER, Davies SR, Friedman PN, Snider JE, et al. PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). NPJ breast cancer. 2016;2:15023.
32. Gendoo DM, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2015;32(7):1097–9.
33. Colbert LE, Petrova AV, Fisher SB, Pantazides BG, Madden MZ, Hardy CW, et al. CHD7 expression predicts survival outcomes in patients with resected pancreatic cancer. Cancer Res. 2014;74(10):2677–87.
34. Patani N, Jiang WG, Newbold RF, Mokbel K. Histone-modifier gene expression profiles are associated with pathological and clinical outcomes in human breast cancer. Anticancer Res. 2011;31(12):4115–25.

35. Laurent B, Ruitu L, Murn J, Hempel K, Ferrao R, Xiang Y, et al. A specific LSD1/KDM1A isoform regulates neuronal differentiation through H3K9 demethylation. Mol Cell. 2015;57(6):957–70.
36. Cayrol C, Lacroix C, Mathe C, Ecochard V, Ceribelli M, Loreau E, et al. The THAP–zinc finger protein THAP1 regulates endothelial cell proliferation through modulation of pRB/E2F cell-cycle target genes. Blood. 2007;109(2): 584–94.
37. Mehta RJ, Jain RK, Leung S, Choo J, Nielsen T, Huntsman D, et al. FOXA1 is an independent prognostic marker for ER-positive breast cancer. Breast Cancer Res Treat. 2012;131(3):881–90.
38. Hicks M, Hu Q, Macrae E, DeWille J. JUNB promotes the survival of Flavopiridol treated human breast cancer cells. Biochem Biophys Res Commun. 2014;450(1):19–24.
39. Kim C, Tang G, Pogue-Geile KL, Costantino JP, Baehner FL, Baker J, et al. Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor–positive breast cancer. J Clin Oncol. 2011;29(31):4160–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.