ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO OPEN SCIENCE FOR OPTIMAL CANCER CARE

ORIGINAL RESEARCH

A three-gene signature marks the time to locoregional recurrence in luminal-like breast cancer

C. Chiodoni[1‡], S. Sangaletti[1‡], M. Lecchi[2‡], C. M. Ciniselli[2], V. Cancila[3], I. Tripodi[1], C. Ratti[1], G. Talarico[1†], S. Brich[4], L. De Cecco[5], P. Baili[6], M. Truffi[7], F. Sottotetti[8], F. Piccotti[7], C. Tripodo[3,9], G. Pruneri[4], T. Triulzi[10], F. Corsi[11,12], V. Cappelletti[13], S. Di Cosimo[13], P. Verderio[2§] & M. P. Colombo[1*§]

[1]Fondazione IRCCS Istituto Nazionale dei Tumori, Experimental Oncology Department, Molecular Immunology Unit, Milan; [2]Fondazione IRCCS Istituto Nazionale dei Tumori, Department of Epidemiology and Data Science, Unit of Bioinformatics and Biostatistics, Milan; [3]University of Palermo School of Medicine, Department of Health Sciences, Tumor Immunology Unit, Palermo; [4]Fondazione IRCCS Istituto Nazionale dei Tumori, Department of Pathology, Milan; [5]Fondazione IRCCS Istituto Nazionale dei Tumori, Experimental Oncology Department, Molecular Mechanisms Unit, Milan; [6]Fondazione IRCCS Istituto Nazionale dei Tumori, Analytical Epidemiology and Health Impact Unit, Milan; [7]Istituti Clinici Scientifici Maugeri IRCCS, Laboratory of Nanomedicine, Pavia; [8]Istituti Clinici Scientifici Maugeri IRCCS, Medical Oncology Unit, Pavia; [9]FIRC Institute of Molecular Oncology (IFOM), Milan; [10]Fondazione IRCCS Istituto Nazionale dei Tumori, Experimental Oncology Department, Molecular Targeting Unit, Milan; [11]Istituti Clinici Scientifici Maugeri IRCCS, Surgery Department, Breast Unit, Pavia; [12]Department of Biomedical and Clinical Sciences 'L. Sacco', University of Milan, Milan; [13]Fondazione IRCCS Istituto Nazionale dei Tumori, Department of Advanced Diagnostics, Biomarkers Unit, Milan, Italy

Available online 30 June 2023

Background: Gene expression profiling (GEP)-based prognostic signatures are being rapidly integrated into clinical decision making for systemic management of breast cancer patients. However, GEP remains relatively underdeveloped for locoregional risk assessment. Yet, locoregional recurrence (LRR), especially early after surgery, is associated with poor survival.

Patients and methods: GEP was carried out on two independent luminal-like breast cancer cohorts of patients developing early ($\leq$5 years after surgery) or late (>5 years) LRR and used, by a training and testing approach, to build a gene signature able to intercept women at risk of developing early LRR. The GEP data of two *in silico* datasets and of a third independent cohort were used to explore its prognostic value.

Results: Analysis of the first two cohorts led to the identification of three genes, *CSTB, CCDC91* and *ITGB1,* whose expression, derived by principal component analysis, generated a three-gene signature significantly associated with early LRR in both cohorts (*P* value <0.001 and 0.005, respectively), overcoming the discriminatory capability of age, hormone receptor status and therapy. Remarkably, the integration of the signature with these clinical variables led to an area under the curve of 0.878 [95% confidence interval (CI) 0.810-0.945]. In *in silico* datasets we found that the three-gene signature retained its association, showing higher values in the early relapsed patients. Moreover, in the third additional cohort, the signature significantly associated with relapse-free survival (hazard ratio 1.56, 95% CI 1.04-2.35).

Conclusions: Our three-gene signature represents a new exploitable tool to aid treatment choice in patients with luminal-like breast cancer at risk of developing early recurrence.

Key words: luminal breast cancer, locoregional recurrence, gene signature, personalized treatment, risk assessment

INTRODUCTION

Breast cancer (BC) is the most frequent malignancy in women, with estrogen receptor (ER)-positive disease representing about 70% of all cases.[1] Luminal-like BC cases are defined by positive tumor immunostaining for ER and/or progesterone receptor (PR) and absence of HER2 overexpression. The values of Ki67 below or above 14% have been considered to split the luminal-like BC into A and B,[2] the latter showing lower expression of ER, PR or estrogen-related genes, higher tumor grade and proliferative index, and a doubled risk of early relapse.[3,4] In light of these features, luminal B-like BCs are usually considered less

*Correspondence to: Dr Mario P. Colombo, Fondazione IRCCS Istituto Nazionale dei Tumori, Experimental Oncology Department, Molecular Immunology Unit, Via Amadeo 42, Milan, 20133, Italy. Tel: +39-02-23902252
E-mail: mariopaolo.colombo@istitutotumori.mi.it (M. P. Colombo).

†Present address: European Institute of Oncology, Laboratory of Hematology-Oncology, Milan, 20141, Italy.

‡CC, SS and ML contributed equally to the work.

§PV and MPC shared the senior authorship.

responsive to endocrine treatments and more sensitive to chemotherapy.[5]

The breakthrough of high-throughput molecular analysis allowed the stratification of BC in the so-called 'intrinsic subtypes' (luminal A, luminal B, HER-2 enriched, basal-like and normal-like), thanks to the pioneering works of Perou et al.[6] and Sorlie et al.[6,7] Since immunohistochemistry (IHC)-based classification and intrinsic molecular subtypes do not overlap, when referring to IHC categorization, the terminology 'luminal-like' should be adopted.[8]

Despite showing clinical features not suggestive of aggressiveness and likelihood of relapse, even in case of luminal A-like BC, there is a small subset of patients undergoing early relapse. Additionally, since luminal-like BC represents three quarters of cases, the number of patients developing a BC event in luminal subtypes double that of all other subtypes. Therefore, proper patient stratification and reliable discrimination between high and low risk of relapse are warranted in luminal-like BC. While currently available methods allow to predict the probability of relapse with some accuracy, there are still many uncertainties due to tumor heterogeneity and mechanisms of recurrence that are not fully explained by ER signaling and cell proliferation. Other factors might be important such as tumor cell dormancy together with microenvironment-related mechanisms including extracellular matrix remodeling and involvement of stromal and immune cells in tumor cell survival.[9,10]

Although molecular profiling is being rapidly integrated into clinical decision making for systemic treatment, its application to locoregional risk assessment is relatively underdeveloped. However, molecular profiling holds the potential to overcome conventional clinico-pathological factors such as primary tumor size, node status and histology and to better assist personalized locoregional risk assessment and management decisions. This would eventually reduce the burden of secondary surgery, which often impairs patient psychological and physical balance, and the known risk of metastasis after the diagnosis of locoregional recurrence (LRR).

This study aims at characterizing luminal-like BC patients at risk of locoregional relapse by gene expression profiling (GEP). For this purpose we analyzed two independent cohorts of cases recurring at different time points after surgery. Using a training and testing set approach, we developed a genomic classifier able to discriminate between early and late recurrent cases and to improve the predictive capability of conventional clinico-pathological features.

MATERIALS AND METHODS

Study patient population

Primary tumor samples were obtained from two retrospective cohorts of luminal-like BC patients developing LRR. The first cohort (cohort A) came from the archival specimen of the Istituti Clinici Maugeri (Pavia, Italy), and the second one (cohort B) from Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, Italy). Specifically, cohort A consisted of a total of 60 consecutive recurrent patients initially diagnosed between 2001 and 2017, and cohort B consisted of a total of 51 consecutive recurrent patients initially diagnosed between 2003 and 2017.

Samples from cohort A referred to the Bruno Boerci Biobank (Pavia, Italy), which preserves and stores tumor specimens exceeding diagnosis from patients undergoing surgery at the Istituti Clinici Maugeri. Samples from cohort B were obtained from tumor specimens exceeding diagnosis preserved in the Pathology Department from the Fondazione IRCCS Istituto Nazionale dei Tumori.

Ethics

Approval for this study was granted by the Ethical Committees of the Istituti Clinici Maugeri and Fondazione IRCCS Istituto Nazionale dei Tumori, specifically with Protocol n. 2590 CE (Pavia, Italy) and INT 0206/16 (Milan, Italy), respectively.

Gene expression analysis

Formalin-fixed paraffin embedded (FFPE) material was revised, to avoid any sign of necrosis and to recover at least 70% of tumor cells. Total RNA was then extracted using the miRNeasy FFPE kit, following the manufacturer's guidelines and automated on QIAcube station (Qiagen, Milan, Italy). Nucleic acid was quantified by Qubit 2.0 Fluorimetric Assay (Thermo Fisher Scientific, Milan, Italy) and the quality was checked by TapeStation4200 (Agilent Technologies, Milan, Italy).

Microarray analysis was carried out with Clariom D Assay, human (ThermoFisher, Milan, Italy) providing information from over 540 000 transcripts. Briefly, the Affymetrix GeneChip WT Pico Kit was used for cDNA preparation, biotin labeling and cRNA synthesis starting from 200 ng of total RNA. The arrays were subsequently incubated for 16 h in an Affymetrix GeneChip 645 hybridization oven and processed with the GeneChip Hybridization. Washing and staining was carried out using the GeneChip HT hybridization, Wash and Stain Kit and with the Affymetrix GeneChip fluidics station 450. The chips were scanned with an Affymetrix GeneChip Scanner 3000 with default manufacturers' settings, and raw data were acquired with the AGCC scan control v4.0. Raw data were normalized using the Signal Space Transformation-Robust Multiarray Analysis (sst-RMA) algorithm implemented in the Transcriptome Analysis Console software (Thermo Fisher, Milan, Italy).

Statistical analysis

Differentially expressed genes were identified in univariate analysis using the nonparametric Kruskal–Wallis test[11] by considering the relapse time, dichotomized as $\leq$5 years compared to >5 years, as outcome measure. Only genes retaining statistical significance after false discovery rate (FDR) adjustment and relevant fold change (FC; i.e. FC $\geq$ 2 and FDR $\leq$ 0.05) were selected (i.e. candidate genes). Candidate genes from cohort A retaining a statistical significance according to the Kruskal–Wallis test and with a relevant FC in cohort B were eventually selected (i.e. confirmed

genes). For explorative purpose the associations between the confirmed genes and the relapse time were further investigated by considering three categorized relapse times: (i) $\leq$2 years versus >5 years and (ii) 2—5 years versus >5 years. A principal component analysis[12] was then implemented to combine the confirmed genes into a score (PCscore) on the training set and then applied to the testing one. To illustrate its potential clinical usefulness, the PCscore's 'optimal' cut-off was computed on the overall cohort by maximizing the Youden index. The relationship between clinico-pathological variables and dichotomized relapse time was investigated by univariate and multivariate logistic regression models. Variables that were statistically significant ($\alpha = 0.05$) in univariate analysis were considered in the clinical multivariate model. A bioclinical model was then obtained by adding the PCscore to the clinical one. For each model, the predictive capability was calculated as the area under the receiver operating characteristic (ROC) curve (AUC) and its corresponding 95% confidence interval (95% CI).

The nonparametric approach of DeLong and Clarke-Pearson was used to compare the discriminatory performance of the two models.[13]

Associations between covariates of the models were investigated according to the nature of the variables through the Kruskal—Wallis test or the Spearman's correlation coefficient (ρ_S). The latter was interpreted according to the criteria suggested by Evans.[14]

Relationships between the gene's expression levels obtained by the original GEP and the IHC assay either quantitative PCR (qPCR) were evaluated by estimating the Spearman's correlation coefficient (ρ_S) together with the 95% CI.

The performance of the PCscore was explored on two *in silico* datasets, the GSE6532[15] and the METABRIC,[16] and on the independent dataset (cohort C) available in our institute.[17] The prognostic role of the PCscore was evaluated by resorting to a univariate Cox regression model. After dichotomizing the PCscore through the 'optimal' cut-off, the survival patterns were estimated using the Kaplan—Meier method and the survival curves were compared using the log-rank test.

All statistical analyses were carried out using SAS software (version 9.4.; SAS Institute, Inc., Cary, NC), adopting a nominal significance level of $\alpha = 0.05$.

Data availability

Expression data generated in this study are deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), with the following accession numbers: GSE207847 (cohort A) and GSE208101 (cohort B), both in the SuperSeries GSE208102. Additional clinical data are available upon request to the corresponding author.

RESULTS

Study population

The distribution of clinical and pathological characteristics among patients from the two cohorts are listed in Table 1.

Table 1. Patient and primary tumor characteristics

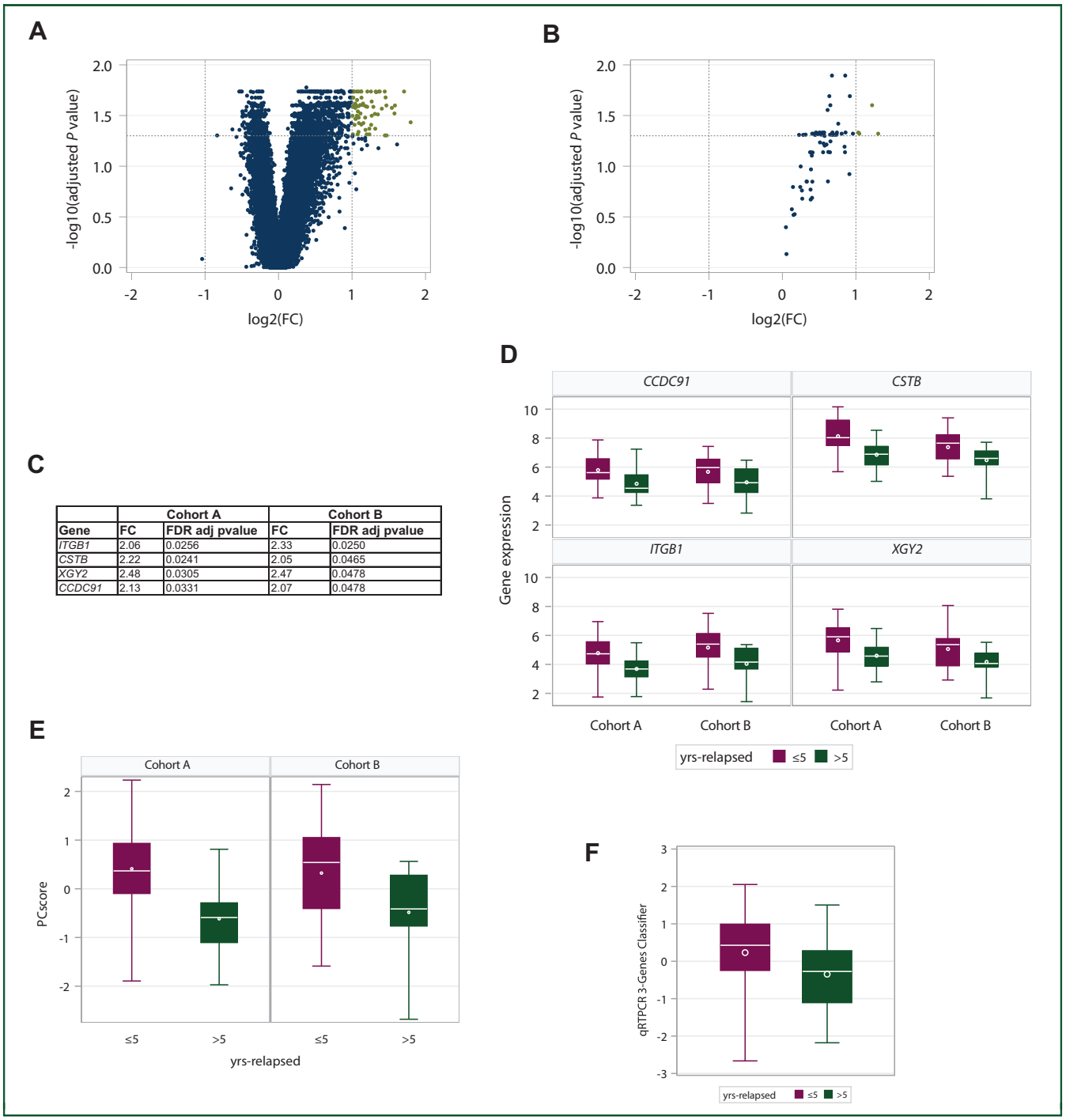
Age at surgery	Cohort A Median (IQR)		Cohort B Median (IQR)	
	62.5 (47-75.5)		55.5 (46-74)	
	Frequency (*N*)	Percentage (%)	Frequency (*N*)	Percentage (%)
pT				
T_1	39	65.00	37	72.55
$T_{\geq 2}$	16	26.67	10	19.61
Missing	5	8.33	4	7.84
pN				
Neg (N0)	33	55.00	24	47.06
Pos (≥N1)	25	41.67	22	43.14
Missing	2	3.33	5	9.80
ER				
Neg	0	0	0	0
Pos (≥10%)	60	100	51	100
PgR				
Neg	9	15.00	9	17.65
Pos (≥10%)	51	85.00	42	82.35
Ki67				
≤14%	37	61.67	14	27.45
>14%	23	38.33	36	70.59
Missing	0	0	1	1.96
Grade				
G1/G2	47	78.33	37	72.55
G3	12	20.00	13	25.49
Missing	1	1.67	1	1.96
Luminal subtype (IHC)				
Luminal A-like	31	51.67	15	29.41
Luminal B-like	27	45.00	32	62.75
Missing	2	3.33	4	7.84
Chemotherapy				
No	45	75.00	30	58.82
Yes	15	25.00	21	41.18
Endocrine therapy				
No	15	25.00	13	25.49
Yes	45	75.00	38	74.51
Radiotherapy				
No	27	45.00	24	47.06
Yes	33	55.00	27	52.94
Surgery type				
MT	24	40	12	23.53
QU	35	58.33	37	72.55
QU + MT	0	0	1	1.96

ER, estrogen receptor; IHC, immunohistochemistry; IQR, interquartile range; MT, mastectomy; Neg, negative; PgR, progesterone receptor; pN, pathological regional lymph nodes; Pos, positive; pT, pathological tumor; QU, quadrantectomy.

Cohorts A and B included patients with ER-positive BC, developing an LRR. Patients received either adjuvant endocrine therapy, chemotherapy, radiotherapy or a combination of these. The two cohorts were well balanced for all the clinical characteristics considered with the unique exception of Ki67, with cohort B including a higher portion of patients with Ki67 values > 14% (70.6% in cohort B compared to 38.3% in cohort A).

Identification of potential genes to generate a locoregional recurrence gene signature

To identify the most relevant LRR-associated genes, the primary BCs of patients in cohort A were analyzed by comparing cases experiencing relapse before or after 5



	Cohort A		Cohort B	
Gene	FC	FDR adj pvalue	FC	FDR adj pvalue
ITGB1	2.06	0.0256	2.33	0.0250
CSTB	2.22	0.0241	2.05	0.0465
XGY2	2.48	0.0305	2.47	0.0478
CCDC91	2.13	0.0331	2.07	0.0478

Figure 1. Differentially expressed genes between patients with locoregional recurrence within or after 5 years from surgery and association between gene expression levels and relapse time. (A) Volcano plot highlighting the 70 differentially expressed genes in cohort A. (B) Volcano plot of the 70 differentially expressed genes quantified in cohort B. (C) List of four genes differentially expressed in cohort A and B, with the corresponding FC and FDR-adjusted *P* value. (D) Distribution of gene expression levels according to the relapse time (≤5 or >5 years) within cohort for each of the four confirmed genes. Each box indicates the 25th and 75th percentiles. The horizontal line and the circle inside the box indicate the median and the mean, respectively. Whiskers indicate the extreme measured values. (E) Distribution of the validated first principal component score (PCscore), in the training and testing set according to the relapse time (*P* value <0.001 and 0.005, respectively), categorized as ≤5 years versus >5 years. Each box indicates the 25th and 75th percentiles. The horizontal line and the circle inside the box indicate the median and the mean, respectively. Whiskers indicate the extreme measured values. (F) Boxplot reporting the distribution of the qPCR three-gene classifier according to the categorized relapse time (≤5 or >5 years) in the overall cohort (*P* value = 0.003).
FC, fold change; FDR, false discovery rate; qPCR quantitative PCR; yrs, years.

years from surgery. We detected a total of 70 differentially expressed genes (Supplementary Table S1, available at https://doi.org/10.1016/j.esmoop.2023.101590) as depicted in the volcano plot (Figure 1A).

Four of the 70 genes found in cohort A retained their statistical significance also in cohort B (Figure 1B and C). The gene expression distribution of the confirmed genes according to the relapse time in the two considered cohorts was analyzed and showed a very similar trend in the two patient subsets (Figure 1C and D).

Among the four confirmed genes, *CCDC91* codes for a protein involved in the transport of secreted proteins through the Golgi, *CSTB* codes for cystatin B (also stefin B), a cathepsin inhibitor, *ITGB1* gene codes for the integrin beta 1 (also known as CD29), a surface receptor regulating extracellular matrix (ECM)—tumor cell interaction, while *XGY2* is a pseudogene (Supplementary Figure S1, available at https://doi.org/10.1016/j.esmoop.2023.101590, reports the distributions of these confirmed genes according to three categorized relapse times).

Validation of the three-gene signature

A qPCR approach was developed and run on all samples to technically validate the three genes, *CSTB*, *ITGB1* and *CCDC91*, normalized by using the b-actin (*ACT*) gene as housekeeping. Since *XGY2* is a pseudogene, we decided to exclude it from the subsequent analysis. Evaluating the relationship between the qPCR gene data and those of the original GEP, we obtained significant Spearman's correlation coefficients (ρ_S) of 0.789 (95% CI 0.700-0.852), 0.753 (95% CI 0.650-0.827) and 0.654 (95% CI 0.522-0.753) for *CSTB*, *ITGB1* and *CCDC91* genes, respectively (Supplementary Figure S2A, available at https://doi.org/10.1016/j.esmoop.2023.101590). Gene expression data were then confirmed at the protein level by IHC analysis. Representative IHC analysis of the three molecules for early- and late-relapse samples, with high and low expression, respectively, and relationships between the gene expression levels obtained by the original GEP and the IHC assay, for each of the three genes, are shown in Supplementary Figure S2B and C, available at https://doi.org/10.1016/j.esmoop.2023.101590. The three confirmed genes were then combined in a multivariate fashion by principal component analysis. The distribution of the validated first standardized principal component score (PCscore = *CSTB* × 0.587116 + *CCDC91* × 0.546487 + *ITGB1* × 0.597199, after gene standardization), in the training and testing set according to the relapse time (*P* value <0.001 and 0.005, respectively), categorized as ≤5 years versus >5 years, is shown in Figure 1E. The corresponding boxplots considering the three categorized relapse time are shown in Supplementary Figure S3, available at https://doi.org/10.1016/j.esmoop.2023.101590. Moreover, by assessing the association between the first principal component of the three qPCR expression genes and the relapse time on the overall case series (cohort A + cohort B), the significant result was confirmed (*P* value = 0.003) (Figure 1F).

Development of an integrated bioclinical model

By looking at the overall case series, univariate logistic regression analysis showed early relapse time (≤5 years) to be significantly associated with age at surgery and inversely associated with PR status, endocrine therapy and radiotherapy (Supplementary Table S2, available at https://doi.org/10.1016/j.esmoop.2023.101590).

The ROC curve of the multivariate clinical model, built by jointly considering these variables, shows an AUC value equal to 0.777 (95% CI 0.689-0.865) (Figure 2A). Taking into account the univariate analysis of the PCscore, a significant association with the early relapse time (Supplementary Table S2, available at https://doi.org/10.1016/j.esmoop.2023.101590) was observed with an AUC value of 0.780 (95% CI 0.695-0.866). For explorative purpose, we pursue the analysis by assessing the PCscore effect after adjustment for tumor-intrinsic molecular subtypes in a bivariate logistic model including the interaction term. To note, the significant effect of PCscore in discriminating early and late recurrence was maintained (*P* value <0.001) regardless of the subtype (luminal A- or B-like) as confirmed by the nonsignificant interaction term in the full model (*P* value = 0.368) (Supplementary Figure S4, available at https://doi.org/10.1016/j.esmoop.2023.101590). Since no associations were found between the clinical variables included in the model and the PCscore (Supplementary Figure S5, available at https://doi.org/10.1016/j.esmoop.2023.101590), a 'bioclinical' model was built combining the PCscore and the clinical classifier. The performance of this 'bioclinical' model was significantly improved (*P* value = 0.005), reaching an AUC value of 0.878 (95% CI 0.810-0.945) (Figure 2A). Finally, as we found the variables endocrine therapy and age at surgery to be significantly associated (Supplementary Figure S6A, available at https://doi.org/10.1016/j.esmoop.2023.101590), the same models were fitted without the variable 'endocrine therapy' in order to avoid colinearity. As shown in Supplementary Figure S6B, available at https://doi.org/10.1016/j.esmoop.2023.101590, the obtained results were comparable to those reported earlier.

The three-gene signature identifies early recurrent cases among low-risk patients

To investigate the clinical relevance of our signature, we considered the subset of 52 patients, belonging to cohorts A and B, who received only endocrine therapy with or without local radiotherapy, but spared from additional chemotherapy, because they were considered to be at good prognosis. Bar plot of the PCscore for each patient according to the time of relapse shows, interestingly, that a high portion of patients who relapsed early has the highest PCscore, suggesting the possibility of identifying this subgroup of patients in need of further therapies, based on our signature (Figure 2B). Also the predictive performance in terms of AUC of the PCscore was confirmed in this subset of patients, as represented by the ROC curve in Figure 2C.

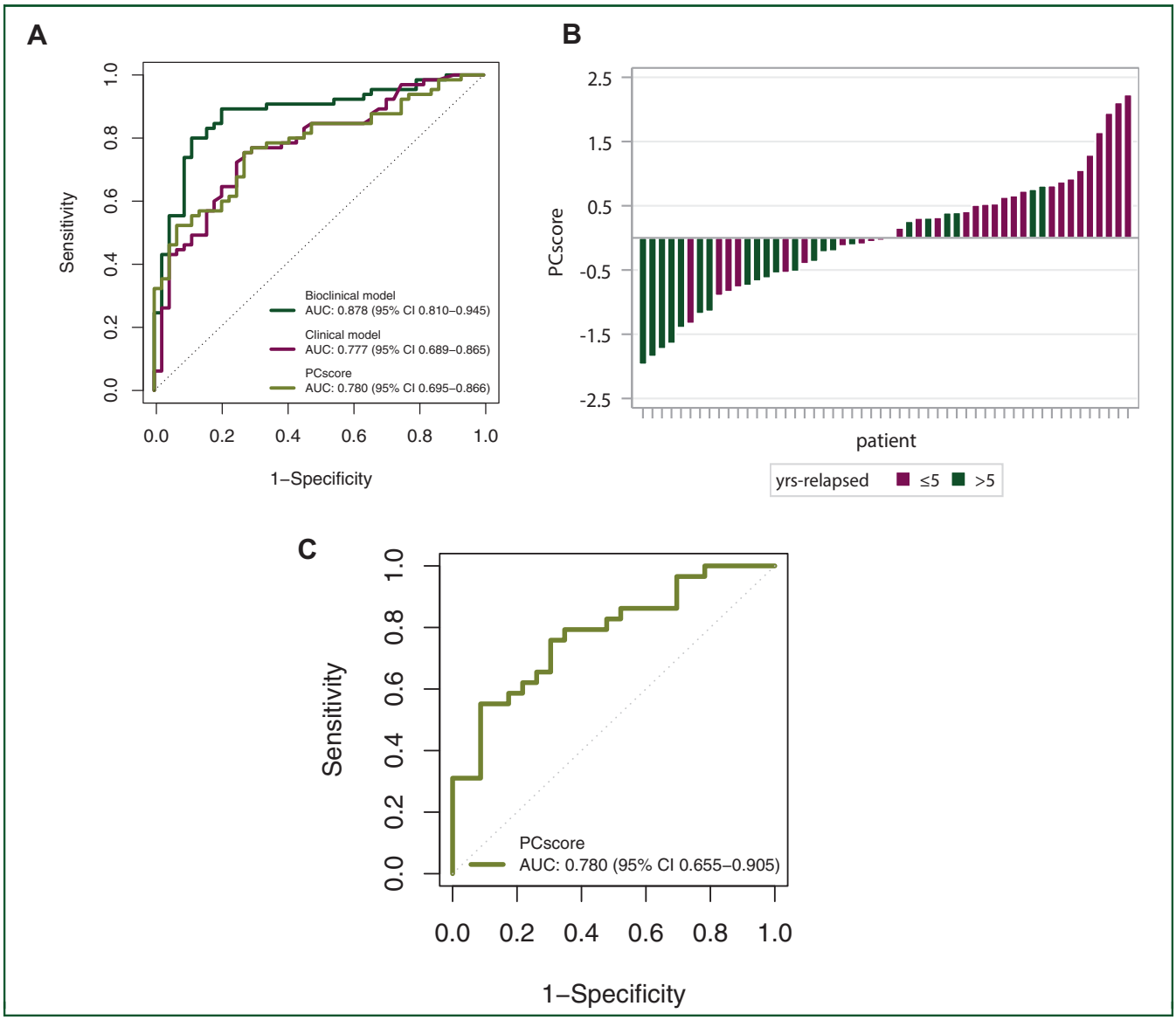


Figure 2. The PCscore improves the predictive role of the clinical model and helps intercepting patients likely to early local relapse among those at 'good' prognosis. (A) ROC curves from PCscore (gray line), clinical (red line) and bioclinical (blue line) models. The performance of the clinical model was significantly improved (*P* value = 0.005) when the PCscore was included, reaching an AUC value of 0.878 (95% CI 0.810-0.945). (B) Bar plot showing the PCscore for each patient according to the relapse time (≤5 or >5 years) in the subset of patients who received only endocrine therapy with or without radiotherapy, but spared from additional chemotherapy and corresponding ROC curve (C) from the PCscore model.
AUC, area under the curve; CI, confidence interval; PCscore, principal component score; pts, patients; ROC, receiver operating characteristic; yrs, years.

Assessing the PCscore in in silico datasets of distant relapse

To extend the clinical value of the PCscore to distant relapse, we considered 512 ER-positive patients from the GSE6532 *in silico* dataset and the subset of 73 ER-positive, HER2-negative, luminal A-like and B-like untreated patients of the METABRIC publicly available dataset. By considering the relapsed patients of each dataset, the PCscore resulted associated with relapse time (≤5 or >5 years), with higher values in the early relapsed patients (Figure 3A and B).

Exploring the prognostic role of the PCscore

To test the PCscore in a prognostic setting and study the association with relapse-free survival (RFS), we used the aforementioned cohorts belonging to the two *in silico* datasets, GSE6532 [hazard ratio (HR) = 1.26; 95% CI 1.08-1.48] and METABRIC (HR = 1.32; 95% CI: 0.91-1.92) and found that patients with low PCscore values showed a better prognosis pattern compared to those with higher values, albeit significant results were obtained only in the GSE6532 dataset (Figure 3C and D).

Additionally, the prognostic value of the PCscore in low-risk BC patients was further exploited in an independent cohort (cohort C) available in our institute of 101 prospectively collected ER-positive, node- and HER2-negative cases from patients receiving only locoregional treatment, with RFS outcome data. Patient and primary tumor characteristics are reported in Supplementary Table S3, available at https://doi.org/10.1016/j.esmoop.2023.101590. At

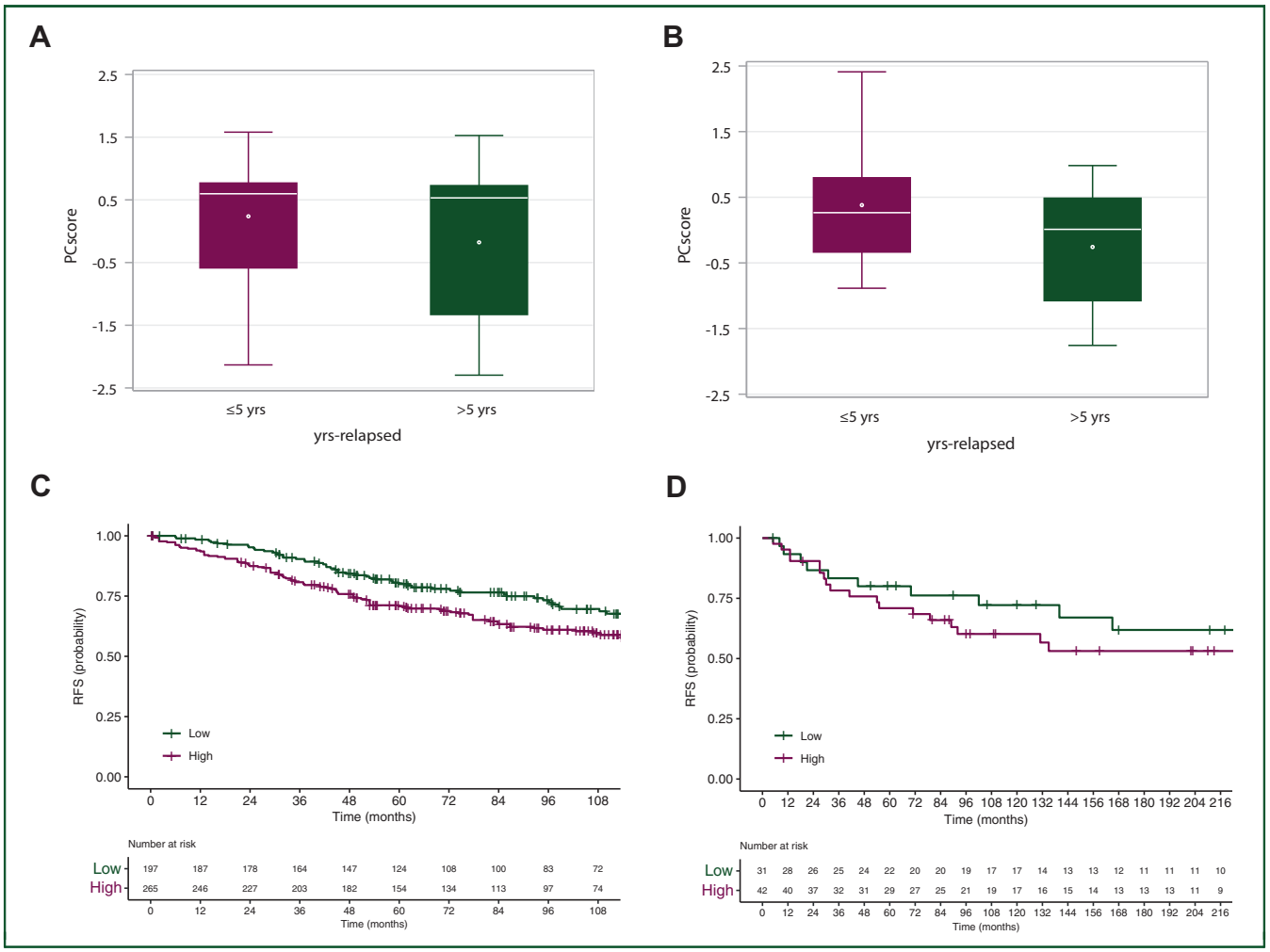


Figure 3. Distribution of the PCscore according to the categorized relapse time and Kaplan–Meier curve according to the dichotomized PCscore. Boxplots reporting the distribution of the PCscore according to the categorized relapse time (≤5 or >5 years) in the GSE6532 (A) and METABRIC (B) *in silico* datasets (two sample *t*-test *P* value = 0.034 and 0.074, respectively). Each box indicates the 25th and 75th percentiles. The horizontal line and the circle inside the box indicate the median and the mean, respectively. Whiskers indicate the extreme measured values. (C and D) RFS curves of the PCscore dichotomized (according to the 'optimal' cut-off = −0.114), in the GSE6532 (log-rank test *P* value = 0.012) and METABRIC (log-rank test *P* value = 0.393) *in silico* dataset, respectively.
PCscore, principal component score; RFS, relapse-free survival; yrs, years.

a median follow-up time of 117 months (interquartile range 89-148 months), the probability of RFS was 64% (95% CI 53% to 73%). The PCscore resulted significantly associated with RFS (HR = 1.56, 95% CI 1.04-2.35), with a decreased probability of RFS by increasing the PCscore (Figure 4A). A similar result was obtained by dichotomizing the PCscore as shown in Figure 4B.

DISCUSSION

A three-gene signature significantly associated with early LRR was developed by gene expression analysis of luminal-like BC primary tumors from two cohorts of recurrent patients by a training and testing approach, and subsequently confirmed in a distinct independent cohort. Combining the molecular information gathered by these genes (defined into a PCscore) with clinico-pathological variables (age at surgery, PgR status, endocrine therapy and radiotherapy) into a 'bioclinical' model, we significantly improved the performance for intercepting patients with luminal-like BC likely to relapse locally within 5 years from surgery (AUC up to 0.878). This suggests that our three-gene signature may detect biological processes not overlapping the clinical variables, improving the chance of identifying patients who are likely to recur early after surgery.

The three-gene signature includes *ITGB1* (integrin b1), *CCDC91* (coiled-coil domain-containing protein 91) and *CSTB1* (cystatin B or stefin B). Integrins are a family of adhesion molecules capable of bidirectional signaling, mediating cell–cell and cell–ECM interactions. After binding to ECM components, integrins trigger intracellular signaling involved in different cellular functions, including proliferation, survival, migration and epithelial-to-mesenchymal transition. Considering that integrins are a key determinant of cellular behavior in response to microenvironmental cues, their deregulation is often associated with cancer development and progression in different tumor types, including BC.[18] The other two genes are far less studied. No data are indeed available for *CCDC91*, which encodes for a protein involved in the regulation of

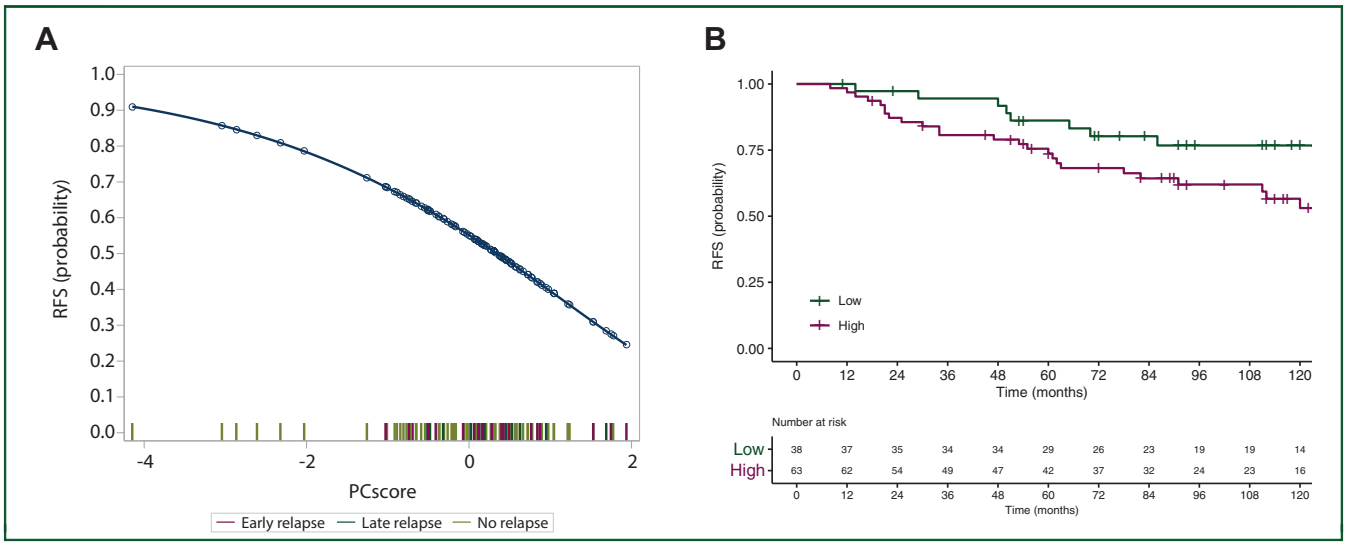


Figure 4. Prognostic role of the PCscore in the independent cohort C. (A) RFS probability curve for the continuous PCscore computed in cohort C. The curve depicts the predicted RFS probability together with the time of relapse in the bottom part. Each segment on the abscissa indicates single patients who relapsed early (in red), late (in blue) or not relapsed (in green). (B) RFS curve of the PCscore dichotomized (according to the 'optimal' cut-off = −0.114) of cohort C (log-rank test *P* value = 0.056).
PCscore, principal component score; RFS, relapse-free survival.

membrane traffic through the trans-Golgi, and only one report in the literature exists for CSTB in BC.[19] Cystatins are a family of endogenous inhibitors of cysteine cathepsins, which are capable of favoring tumor growth and metastasis by activation of the extracellular proteolytic cascades and degradation of the ECM,[20] but are also capable of inducing tumor cell apoptosis.[21] MMTV-PyMT spontaneous model of metastatic BC crossed with CSTB (*Stfb*)-deficient mice showed reduced size of primary tumors but no effect on the rate of metastasis in comparison to their wild-type counterparts. This study on mouse models also showed that stefin B facilitates tumor growth inducing tumor resistance to both oxidative stress and apoptosis.[19]

Although tumor size and nodal status were historically associated with LRR, they were not significant factors in the univariate analysis, suggesting that stage is a predictor of recurrence but not of the timing of recurrence after surgery.[22] As expected, univariate results indicated that the probability of early recurrence was higher in patients not receiving adjuvant therapy. Therefore, our three-gene signature could represent a suitable tool for a more precise risk definition of developing LRR in women affected with BC.[23] Its application should help in deciding, for each patient, the recommendation for adjuvant therapy and its intensity. Indeed, the ability to predict the timing of recurrence has a considerable impact on patient management, as it could drive treatment escalation for patients at risk of early recurrence. Hence, our three-gene signature holds the potential to identify patients at high risk for LRR, despite surgery and radiotherapy, who may benefit from radiation dose escalation or combinations with systemic treatments, as well as to refine selection criteria for the use of partial breast irradiation or whole-breast hypofractionated accelerated treatments.[24-26]

It should also be noted that our results from a multivariable analysis demonstrate that systemic therapy is able to reduce the risk of LRR. It follows that the differences in LRR among cases with different levels of the three-gene signature might be modified in a population receiving more extensive systemic therapy and, therefore, might guide the decision of the adjuvant treatment.

The clinical value of the three-gene signature may be also extended to distant relapse as indicated by *in silico* analysis of the subset of ER-positive relapsed patients from two publicly available datasets in which the score was significantly associated with relapse time, with higher values in the early relapsed patients, and a better prognosis pattern in patients showing low three-gene score values compared to those with higher values was observed. However, we are aware that the identified cut-off should be validated in further *ad hoc* studies implemented on independent cohorts.

A major limitation of our study cohorts was the lack of a luminal-like BC patient population not developing LRR for evaluation the three-gene signature specificity. However, in the additional dataset that included 101 ER-positive, node- and HER2-negative patients with or without LRR, patients without relapse showed a lower three-gene score than relapsed patients, suggesting that the tumor expression levels of *ITGB1*, *CCDC91* and *CSTB* are crucial in promoting luminal-like BC progression. In support of this, the three-gene signature, as a continuous variable, was found significantly associated with RFS (HR = 1.56, 95% CI 1.04-2.35).

Altogether, although our findings need to be further confirmed, the three-gene signature characterization of each patient's tumor holds the potential to help in the selection of the adjuvant treatment in luminal-like BCs, which is a critical challenge for public health, in terms of patient quality of life and survival, reduction in (re)surgery and hospitalization, as well as in the psychological effects that cancer recurrences present to the patients.

ACKNOWLEDGEMENTS

The authors acknowledge the Integrated Biology Platform of the Fondazione IRCCS Istituto Nazionale dei Tumori for running the gene expression profiles.

FUNDING

This work was supported by AIRC (Associazione Italiana Ricerca sul Cancro), grants to MPC [grant numbers AIRC IG 2016 #18425, AIRC IG 2020 #24363] and Italian Ministry of Health, 'Ricerca corrente' funds.

DISCLOSURE

The authors have declared no conflicts of interest.

REFERENCES

1. Giaquinto AN, Sung H, Miller KD, et al. Breast Cancer Statistics, 2022. *CA Cancer J Clin*. 2022;72(6):524-541.
2. Cheang MCU, Chia SK, Voduc D, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst*. 2009;101(10):736-750.
3. Wirapati P, Sotiriou C, Kunkel S, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*. 2008;10(4):R65.
4. Jatoi I, Anderson WF, Jeong JH, Redmond CK. Breast cancer adjuvant therapy: time to consider its time-dependent effects. *J Clin Oncol*. 2011;29(29):3948-3948.
5. Ignatiadis M, Sotiriou C. Luminal breast cancer: from biology to treatment. *Nat Rev Clin Oncol*. 2013;10(9):494-506.
6. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752.
7. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98(19):10869-10874.
8. Schettini F, Braso-Maristany F, Kuderer NM, et al. A perspective on the development and lack of interchangeability of the breast cancer intrinsic subtypes. *NPJ Breast Cancer*. 2022;8(1):85.
9. Hanker AB, Sudhan DR, Arteaga CL. Overcoming endocrine resistance in breast cancer. *Cancer Cell*. 2020;37(4):496-513.
10. Saatci O, Huynh-Dam KT, Sahin O. Endocrine resistance in breast cancer: from molecular mechanisms to therapeutic strategies. *J Mol Med (Berl)*. 2021;99(12):1691-1710.
11. Hollander M, Wolfe DA, Chicken E. *Nonparametric Statistical Methods*. 2nd ed. New York: John Wiley & Sons; 1999.
12. Joliffe I. *Principal Component Analysis*. New York: Springer-Verlag; 1986.
13. Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves - a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
14. Evans JD. *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove: Thomson Brooks/Cole Publishing Co; 1996.
15. Loi S, Haibe-Kains B, Desmedt C, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol*. 2007;25(24):3790-3790.
16. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-352.
17. Callari M, Musella V, Di Buduo E, et al. Subtype-dependent prognostic relevance of an interferon-induced pathway metagene in node-negative breast cancer. *Mol Oncol*. 2014;8(7):1278-1289.
18. Yousefi H, Vatanmakanian M, Mahdiannasser M, et al. Understanding the role of integrins in breast cancer invasion, metastasis, angiogenesis, and drug resistance. *Oncogene*. 2021;40(6):1043-1063.
19. Butinar M, Prebanda MT, Rajkovic J, et al. Stefin B deficiency reduces tumor growth via sensitization of tumor cells to oxidative stress in a breast cancer model. *Oncogene*. 2014;33(26):3392-3400.
20. Joyce JA, Baruch A, Chehade K, et al. Cathepsin cysteine proteases are effectors of invasive growth and angiogenesis during multistage tumorigenesis. *Cancer Cell*. 2004;5(5):443-453.
21. Droga-Mazovec G, Bojic L, Petelin A, et al. Cysteine cathepsins trigger caspase-dependent cell death through cleavage of Bid and antiapoptotic Bcl-2 homologues. *J Biol Chem*. 2008;283(27):19140-19150.
22. Ustaalioglu BBO, Balvan O, Bilici A, et al. The differences of clinicopathological factors for breast cancer in respect to time of recurrence and effect on recurrence-free survival. *Clin Transl Oncol*. 2015;17(11):895-902.
23. Haffty BG, Buchholz TA. Molecular predictors of locoregional recurrence in breast cancer: ready for prime time? *J Clin Oncol*. 2010;28(10):1627-1629.
24. Leonardi MC, Scognamiglio IR, Maisonneuve P, et al. Mastectomy alone for pT1-2 pN0-1 breast cancer patients: when postmastectomy radiotherapy is indicated. *Breast Cancer Res Treat*. 2021;188(2):511-524.
25. Witt JS, Wisinski KB, Anderson BM. Concurrent radiation and modern systemic therapies for breast cancer: an ever-expanding frontier. *Clin Breast Cancer*. 2021;21(2):120-127.
26. Dong Y, Zhang WW, Wang J, et al. The 21-gene recurrence score and effects of adjuvant radiotherapy after breast conserving surgery in early-stage breast cancer. *Future Oncol*. 2019;15(14):1629-1639.