A 17-gene expression-based prognostic signature associated with the prognosis of patients with breast cancer
A STROBE-compliant study

Jin-Xian Qian, MD[a],[*], Min Yu, MM[b], Zhe Sun, MM[a], Ai-Mei Jiang, MM[a], Bo Long, MD[c],[*]

Abstract

Identification of reliable predictive biomarkers for patients with breast cancer (BC).

Univariate Cox proportional hazards regression model was conducted to identify genes correlated with the overall survival (OS) of patients in the TCGA-BRCA cohort. Functional enrichment analysis was conducted to investigate the biological meaning of these survival related genes. Then, patients in TCGA-BCRA were randomly divided into training set and test. Least absolute shrinkage and selection operator (LASSO) penalized Cox regression model was performed and the risk score of BC patients in this model was used to build a prognostic signature. The prognostic performance of the signature was evaluated in the training set, test set, and an independent validation set GSE7390.

2519 genes were demonstrated to be significantly associated with the OS of BC patients. Functional annotation of the 2519 genes suggested that these genes were associated with immune response and protein synthesis related gene ontology terms and pathways. 17 genes were identified in the LASSO Cox regression model and used to construct a 17-gene signature. Patients in the 17-gene signature low risk group have better OS and event-free survival compared with those in the 17-gene signature high risk group in the TCGA-BRCA cohort. The prognostic role of the 17-gene signature has been confirmed in the validation cohort. Multivariable Cox proportional hazards regression model suggested the 17-gene signature was an independent prognostic factor in BC.

The 17-gene signature we developed could successfully classify patients into high- and low-risk groups, indicating that it might serve as candidate biomarker in BC.

Abbreviations: BC = breast cancer, CoxPH = Cox proportional hazards regression model, EFS = event-free survival, ER = estrogen receptor, GO = gene ontology, GSEA = gene set enrichment analysis, HER2 = human epidermal growth factor receptor 2, KEGG = Kyoto Encyclopedia of Genes and Genomes, KM = Kaplan–Meier, LASSO = least absolute shrinkage and selection operator, OS = overall survival, PR = progesterone receptor, ROC = receiver operating characteristic curve.

Keywords: breast cancer, least absolute shrinkage and selection operator, prognosis significance, prognostic signature

1. Introduction

Breast cancer (BC), the second most frequent malignance in females with an estimated 1,676,000 newly diagnosed cases annually, represents a common presentation worldwide but is of special concern in developing countries with limited screening policies.[1] Radical mastectomy is considered frontline treatment for early-stage BC patients[2,3]; Adjuvant treatment (chemother-

apy, radiation therapy, hormone therapy, biotherapy, etc) of BC is designed to treat micrometastatic disease.[4,5] Owning to the fact that many patients are diagnosed with advanced BC, which makes them lose the chance of surgical treatment and the disease remains incurable.

Currently, conventional histopathological test is considered the most common and reliable method for the prognosis prediction and treatment decision in patients with BC.[6] However, it often fails to explain the diversity of genetic burden experienced by the independent individuals and to classify patients into different risk groups.[6] With the development of high-throughput sequencing, several molecular, including SPAG9,[7] neogenin,[8] thioredoxin1 (Trx1),[9] and AGR3,[10] have been severed as biomarkers that are correlated with clinical stage, diagnostic subtype, and prognosis of patients with BC. Nevertheless, few biomarkers have been effectively applied to clinical settings. Least absolute shrinkage and selection operator (LASSO) is a regression analysis method for simultaneous feature selection and regularization,[11] which has been used to screen a variety of tumor-associated biomarkers.[12,13]

Herein, we retrospectively analyzed the gene expression profiles of 1080 BC patients and developed a 17-gene based signature, demonstrating that the 17-gene signature might be a candidate prognostication factor in patients with BC.

2. Methods and materials

2.1. BC gene expression studies

BC gene expression profile was obtained from TCGA-BRCA[14] cohort and GSE7390.[15] In the TCGA-BRCA cohort, we included primary BC patients with survival information documented, and we treated this BC cohort as training set and test set. GSE7390, an Affymetrix Human Genome U133A Array, included 198 gene expression profiles of 198 patients and was treated as an independent validation cohort in this study. Ethical approval is not necessary for this study.

2.2. Model training and signature construction

Univariate Cox proportional hazards regression model (CoxPH) was conducted to identify genes relating with the overall survival (OS) of BC patients in the TCGA-BRCA cohort, and then the TCGA-BRCA cohort was randomly classified into tow cohorts (training set and test set) in a 2:3 ratio. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were conducted to figure out the biological significance of these survival related genes using DAVID (v6.8).[16] Subsequently, we performed LASSO penalized CoxPH model in the training set by using the R package "glmnet."[17] Genes with coefficient not shrunk exactly to zero were applied to form a multigene based prognostication signature. The risk score of each BC patients was calculated based on the coefficients of each gene in the LASSO penalized CoxPH model.

2.3. Evaluation of the prognostic performance of the 17-gene signature

According to the cutoff value derived from time-dependent receiver operating characteristic curve (ROC) analysis, BC patients were categorized into 2 risk groups (17-gene signature high risk group and the 17-gene signature low risk group).[18] We compared the OS and event-free survival (EFS) of high risk and

low risk BC patients in the training set, test set, and validation set (GSE7390) using log-rank test and Kaplan–Meier (KM) curves. The risk score of each BC patients in the validation set was estimated according to the coefficients of the 17 genes in the LASSO penalized CoxPH in the training set. Univariate and multivariable CoxPH models were also conducted.

2.4. Gene set enrichment analysis (GSEA)

Finally, to recognize related mechanisms that the 17-gene signature influenced on the survival of BC patients, we conducted GSEA[19] in the training set and test based on the risk score of each BC patients. Patients were assigned to 2 different risk groups as introduced above.

2.5. Statistical analysis

All the statistical analyses in the present study were performed using R 3.5.2. For the survival analysis including Kaplan–Meier curve, univariate CoxPH model and multivariable CoxPH model, P values less than .05 were considered statistically significant. For the GSEA analysis, gene sets with normal P value less than .05 and false discovery rate less than 25% were considered significantly enriched.

3. Results

3.1. Characteristics of patients with BC in the 3 experimental cohorts

The TCGA-BRCA cohort included a total of 1248 samples, 1101 of which are BC, and among the 1101 BC samples, the OS information is available. After randomization, 648 cases and 432 cases were categorized into the training set and the test set, respectively (the baseline features of patients in the training set and test set were shown in supplementary Table 1, http://links. lww.com/MD/E8 and supplementary Table 2, http://links.lww. com/MD/E9). The validation set consisted of a total of 199 BC patients (median age [range], 46 [24–60] years) was included in (the baseline features of BC patients in the validation set was summarized in supplementary Table 3, http://links.lww.com/ MD/E11).

3.2. Construction of the 17-gene signature

A total of 2519 genes were demonstrated to be significantly related with the OS of BC patients using univariate CoxPH model. The KEGG pathway and GO enrichment analysis suggested that the 2519 gene were mostly enriched in immune related pathways (antigen processing and presentation, primary immunodeficiency, allograft rejection, and graft-versus-host disease, Fig. 1A) and GO terms (nuclear-transcribed mRNA catabolic process, SRP-dependent cotranslational protein targeting to membrane, viral transcription, nonsense-mediated decay, translational initiation, translation, inflammatory response, T cell receptor signaling pathway, rRNA processing, adaptive immune response, immune response, interferon-gamma-mediated signaling pathway, regulation of immune response, and T cell costimulation, Fig. 1B), indicating that these genes were mostly related with immune response and protein synthesis. Thus, the 2519 genes were applied in a LASSO penalized CoxPH model, and 17 genes were identified after feature selection in this model. Therefore, we formed a 17-gene signature according to the risk
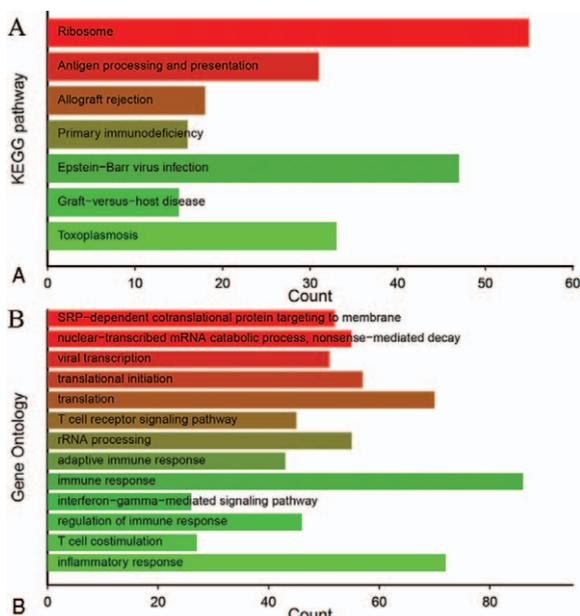
Figure 1. The details of the 17-gene prognostic signature. (A) The risk score of breast cancer patients calculated based on the LASSO penalized Cox proportional hazards regression model in the 17-gene signature low risk group and 17-gene signature high risk group. (B) The survival status and time of breast cancer patients in the 17-gene signature low risk group and 17-gene signature high risk group. (C) The expression levels of the 17 genes in the 17-gene signature low risk group and 17-gene signature high risk group.

score of each BC patients (supplementary Table 4, http://links.lww.com/MD/E12 and Fig. 2).

3.3. The prognostication performance of the 17-gene signature

The results of KM curves indicated that patients in the 17-gene signature low risk group were associated superior OS than those in the 17-gene signature high risk group in the training set (univariate CoxPH, HR = 94.06003, 95% CI: 33.54986–263.7057; log-rank $P = 5.67E-18$, supplementary Table 5, http://links.lww.com/MD/E13 and Fig. 3A) and test set (univariate CoxPH, HR = 14.71657, 95% CI: 3.273202–66.16684; log-rank $P = .000455$, supplementary Table 6, http://links.lww.com/MD/E14 and Fig. 3B). For the EFS, patients in the 17-gene signature low risk group were associated with better EFS compared with those in the 17-gene signature high risk group in the training set (univariate Cox regression model, HR = 6.768481, 95% CI: 1.563461–29.30188; log-rank $P = .010537$, supplementary Table 7, http://links.lww.com/MD/E15 and Fig. 3C) and test set (univariate Cox regression model, HR = 6.052806, 95% CI:1.321835–27.71637; log-rank $P = .020374$, supplementary Table 8, http://links.lww.com/MD/E16 and Fig. 3D). Furthermore, we analyzed the prediction value of the 17-gene signature in GSE7390. As shown in Figure 4, the OS of patients in the 17-gene signature low risk group were significantly longer as compared to that in the 17-gene signature high risk group (log-rank $P = .047$). Furthermore, we also divided BC patients in the TCGA-BRCA (including the training set and test set) into triple negative BC (TNBC) and non-TNBC, and
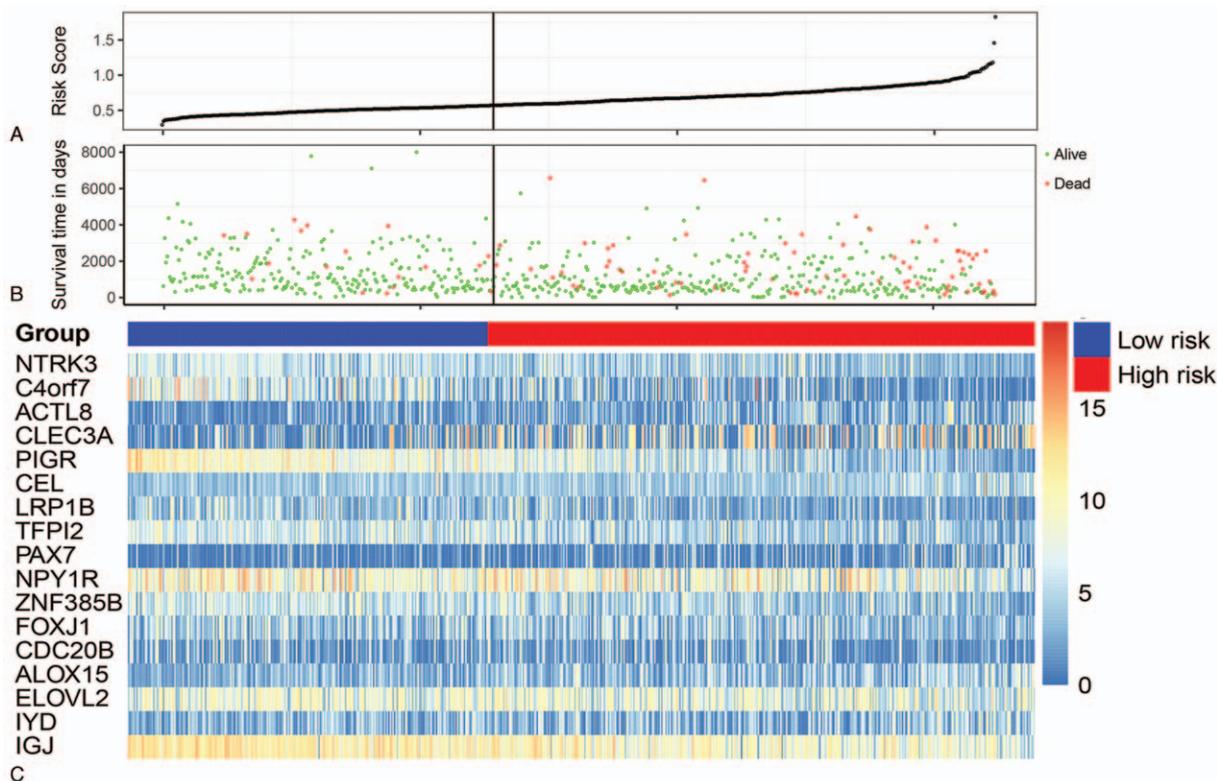
Figure 2. Functional annotation of the 2519 genes correlated with the overall survival of breast cancer patients. (A) KEGG signaling pathways that the 2519 genes were enriched with. (B) GO terms that the 2519 genes were enriched. The more significant the P value of the GO terms and KEGG pathways, the closer the color of the corresponding bar is to red. The count of the GO terms and KEGG pathways means the number of genes enriched in the corresponding terms.
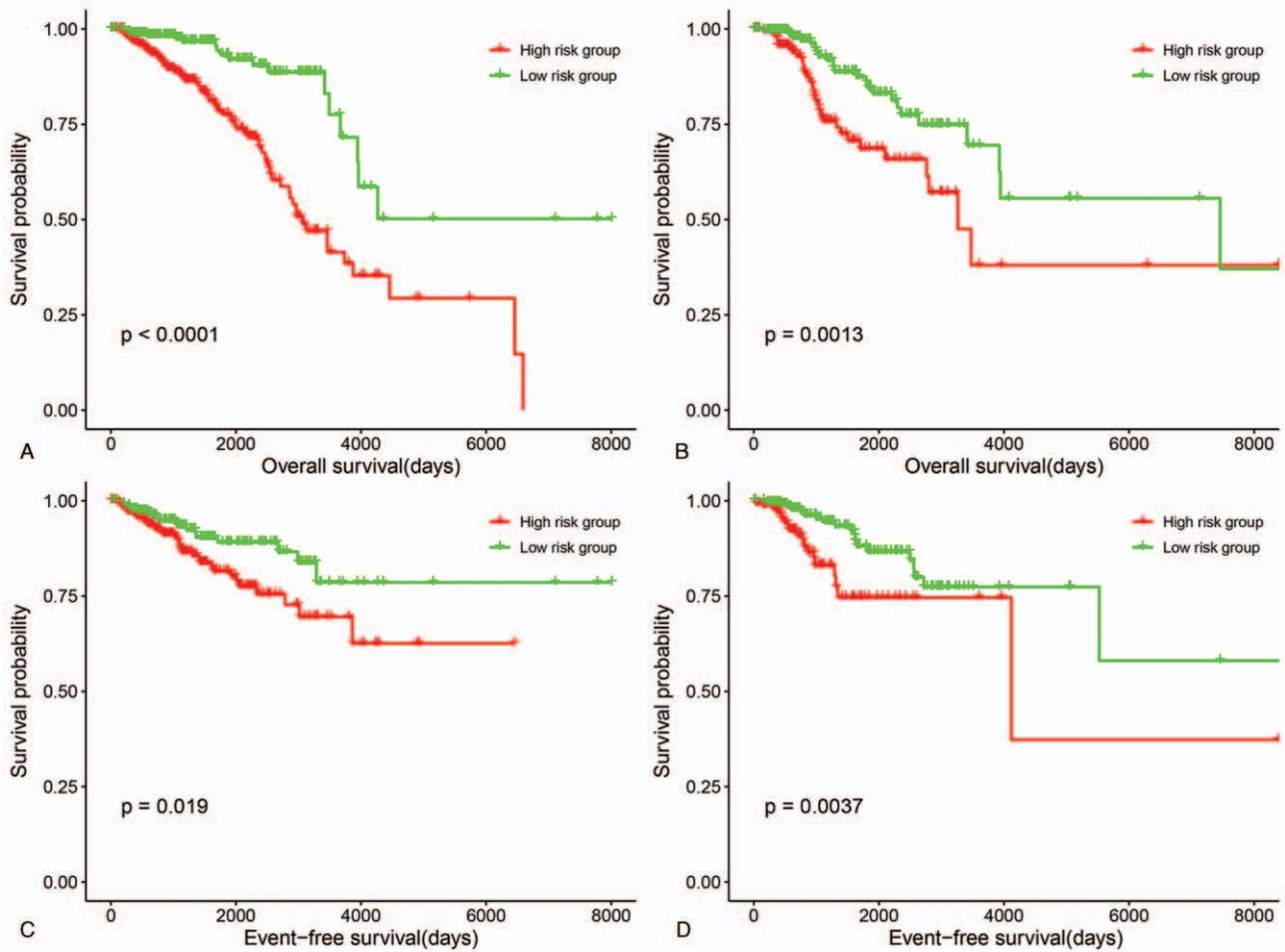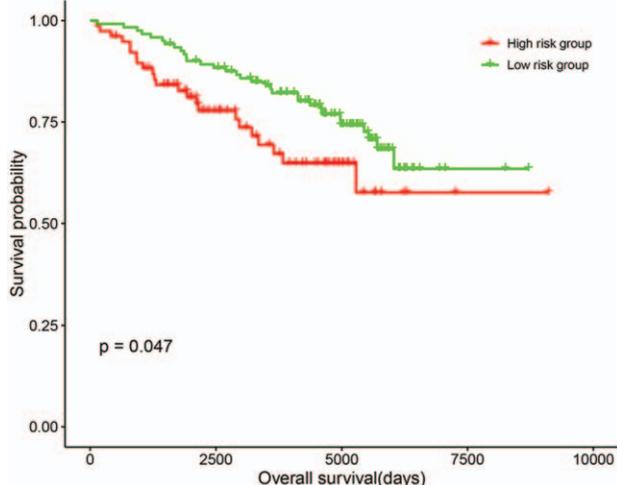
Figure 3. The influence of the 17-gene signature on the overall survival and event-free survival in the training set and test set. (A) Overall survival in the training set. (B) Overall survival in the test set. (C) Event-free survival in training set. (D) Event-free survival in the test set.

investigated the survival relevance of the 17-gene signature of patients in the 2 subgroups. As shown in supplementary Figure 1E–H, http://links.lww.com/MD/E7 the 17-gene signature could significantly stratify the both TNBC and non-TNBC

Figure 4. The influence of the 17-gene signature on the overall survival in the validation set.

patients into different survival groups in the training set and test set. In the validation set GSE7390, the pathological stage and entire statuses of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) were not reported, thus, we could not perform subgroup analysis in GSE7390 the same with in the training set and test. However, the ER status patients in GSE7390 were documented, so we divided the patients into ER-positive group and ER-negative group accordingly and investigated the survival differences of patients in the 17-gene signature low risk group and 17-gene signature high risk group in the 2 subgroups. As shown in supplementary Figure 2, http://links.lww.com/MD/E10 both ER-positive and ER-negative patients in the 17-gene signature low risk group showed better overall survival than those in the 17-gene signature high risk group. The above results indicated that the 17-gene signature had significant prognostication capability in patients with BC.

3.4. Results of GSEA

Finally, in order to investigate the biological foundation that the 17-gene signature affect the survival of BC patients and lay the foundation of future study, we performed GSEA in the training set and test set. As shown in supplementary Table 9, http://links.lww.com/MD/E17 BC sample in the 17-gene high risk group was

significantly associated with unfolded protein response, mTORC1 signaling, MYC signaling pathway, E2F signaling pathway, and G2M checkpoint.

4. Discussions

As mentioned above, BC represents a type of malignant disease with high incidence, limited screening, diagnosis and management approach, and poor prognosis. Therefore, the discovery of novel tumor markers that could predict the survival of patients is of great significance for the treatment and prognosis of patients with BC.

In this study, we identified 2519 genes linked to the OS of patients with BC, and the result of functional annotation of the 2519 genes indicated that they participated in immune response and protein synthesis. Thus, we built a 17-gene (NTRK3, C4orf7, ACTL8, CLEC3A, PIGR, CEL, LRP1B, TFPI2, PAX7, NPY1R, ZNF385B, FOXJ1, CDC20B, ALOX15, ELOVL2, IYD, and IGJ) based prognostic signature in BC. Moreover, we assessed the prognostic performance of the 17-gene signature 3 different cohorts (training, test, and validation set) and the results suggested that the 17-gene signature might be a candidate prognostication factor in patients with BC.

Compared with the conventional biomarkers in BC, including ER, PR, and HER2, our 17-gene signature has better clinical applicability, and it could be well applied to the clinic only by immunohistochemical method. Meanwhile, among the 17 genes, 12 genes were reported in the carcinogenesis and progression of BC including NTRK3,[20–22] ACTL8,[23] CLEC3A,[24] LRP1B,[25] TFPI2,[26] PAX7,[27] NPY1R,[28,29] ZNF385B,[30] FOXJ1,[31] CDC20B,[32] ALOX15,[33] and ELOVL2,[34,35] indicating that the 17-gene based prognostic signature is reliable.

The results of GSEA suggested, the 17-gene signature might affect the survival of BC patients through unfolded protein response, mTORC1 signaling, MYC signaling pathway, E2F signaling pathway, and G2M checkpoint. Actually, the above mechanisms had been found in BC. Shajahan-Haq et al demonstrated that MYC regulated the unfolded protein response in endocrine resistant BC,[36] and Notte et al suggested that unfolded protein response was correlated with the Taxol-induced autophagy and apoptosis in BC cells.[37] Davis et al, Guichard et al, and Lu et al demonstrated that mTORC1 signaling was dysregulated in BC, which provided the possibility of treatment for BC.[38–40] Rennhack et al showed that E2F signaling mediated metastasis HER2 positive BC patients.[41] Kawamoto et al demonstrated that cyclin B1 and CDC28A were differently expressed between normal breast and BC, which could be applied to differentiate between precancerous human breast lesions and advanced BC.[42] Thus, the results of GSEA gave us a cleaner and deeper understanding of the effect of the 17-gene signature in BC and laid the foundation for future studies.

Our study has several limitations. Firstly, the mechanisms that the 17-gene signature affected the survival of BC are derived from statistical inference with no in vivo and in vitro experiment validation. We plan to further explore such mechanisms in vivo and in vitro in future studies. Secondly, the prognostic performance the 17-gene signature has not been confirmed in clinical practice. To the end, our following research will emphasize testing the performance of the 17-gene in clinical trials.

In summary, the 17-gene signature we suggested was effective to categorize BC patients into different survival groups, and it might be treated as candidate prognostication factor in clinical settings.

Author contributions

Conceptualization: Bo Long.
Data curation: Jin-Xian Qian, Min Yu, Bo Long.
Formal analysis: Min Yu, Ai-Mei Jiang.
Investigation: Jin-Xian Qian, Min Yu.
Methodology: Jin-Xian Qian.
Resources: Zhe Sun.
Software: Jin-Xian Qian, Min Yu.
Validation: Zhe Sun, Ai-Mei Jiang.
Visualization: Zhe Sun.
Writing – original draft: Jin-Xian Qian.
Writing – review & editing: Ai-Mei Jiang, Bo Long.

References

[1] Cedolini C, Bertozzi S, Londero AP, et al. Type of breast cancer diagnosis, screening, and survival. Clin Breast Cancer 2014;14:235–40.
[2] de la Mare JA, Contu L, Hunter MC, et al. Breast cancer: current developments in molecular approaches to diagnosis and treatment. Recent Pat Anticancer Drug Discov 2014;9:153–75.
[3] Peart O. Breast intervention and breast cancer treatment options. Radiol Technol 2015;86:535M–58M. quiz 559–562.
[4] Loibl S, Denkert C, von Minckwitz G. Neoadjuvant treatment of breast cancer – clinical and research perspective. Breast 2015;24(Suppl 2):S73–7.
[5] Freedman RA. Treatment of breast cancer in the elderly. Curr Oncol Rep 2015;17:51.
[6] Viale G. Specifics of breast cancer diagnosis in everyday practice. Arkh Patol 2013;75:54–61.
[7] Yang C, Shen B, Zhang J, et al. Sperm-associated antigen 9 over-expression correlates with poor prognosis and insensitive to Taxol treatment in breast cancer. Biomarkers 2016;21:62–7.
[8] Xing W, Li Q, Cao R, et al. Neogenin expression is inversely associated with breast cancer grade in ex vivo. World J Surg Oncol 2014;12:352.
[9] Park BJ, Cha MK, Kim IH. Thioredoxin 1 as a serum marker for breast cancer and its use in combination with CEA or CA15-3 for improving the sensitivity of breast cancer diagnoses. BMC Res Notes 2014;7:7.
[10] Garczyk S, von Stillfried S, Antonopoulos W, et al. AGR3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection. PLoS One 2015;10:e0122106.
[11] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc B 2011;73:273–82.
[12] Liu XP, Yin XH, Meng XY, et al. Development and validation of a 9-gene prognostic signature in patients with multiple myeloma. Front Oncol 2018;8:615.
[13] Song J, Xu Q, Zhang H, et al. Five key lncRNAs considered as prognostic targets for predicting pancreatic ductal adenocarcinoma. J Cell Biochem 2018;119:4559–69.
[14] Cancer Genome Atlas NComprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
[15] Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res 2007;13:3207–14.
[16] Huang Da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1–3.
[17] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
[18] Lorent M, Giral M, Foucher Y. Net time-dependent ROC curves: a solution for evaluating the accuracy of a marker to predict disease-related mortality. Stat Med 2014;33:2379–89.
[19] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
[20] Tognon CE, Somasiri AM, Evdokimova VE, et al. ETV6-NTRK3-mediated breast epithelial cell transformation is blocked by targeting the IGF1R signaling pathway. Cancer Res 2011;71:1060–70.
[21] Del Castillo M, Chibon F, Arnould L, et al. Secretory breast carcinoma: a histopathologic and genomic spectrum characterized by a joint specific ETV6-NTRK3 gene fusion. Am J Surg Pathol 2015;39:1458–67.

[22] Tognon C, Knezevich SR, Huntsman D, et al. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell 2002;2:367–76.

[23] Yao J, Caballero OL, Yung WK, et al. Tumor subtype-specific cancer-testis antigens as potential biomarkers and immunotherapeutic targets for cancers. Cancer Immunol Res 2014;2:371–9.

[24] Ni J, Peng Y, Yang FL, et al. Overexpression of CLEC3A promotes tumor progression and poor prognosis in breast invasive ductal cancer. Onco Targets Ther 2018;11:3303–12.

[25] Wang Y, Guo L, Feng L, et al. Single nucleotide variant profiles of viable single circulating tumour cells reveal CTC behaviours in breast cancer. Oncol Rep 2018;39:2147–59.

[26] Xu C, Wang H, He H, et al. Low expression of TFPI-2 associated with poor survival outcome in patients with breast cancer. BMC Cancer 2013;13:118.

[27] Wang R, Bhat-Nakshatri P, Padua MB, et al. Pharmacological dual inhibition of tumor and tumor-induced functional limitations in a transgenic model of breast cancer. Mol Cancer Ther 2017;16: 2747–58.

[28] Guerin B, Dumulon-Perreault V, Tremblay MC, et al. [Lys(DOTA)4] BVD15, a novel and potent neuropeptide Y analog designed for Y1 receptor-targeted breast tumor imaging. Bioorg Med Chem Lett 2010;20:950–3.

[29] Liu L, Xu Q, Cheng L, et al. NPY1R is a novel peripheral blood marker predictive of metastasis and prognosis in breast cancer patients. Oncol Lett 2015;9:891–6.

[30] Walker LC, Marquart L, Pearson JF, et al. Evaluation of copy-number variants as modifiers of breast and ovarian cancer risk for BRCA1 pathogenic variant carriers. Eur J Hum Genet 2017;25:432–8.

[31] Zhou Q, Jackson-Cook C, Lyon D, et al. Identifying molecular features associated with psychoneurological symptoms in women with breast cancer using multivariate mixed models. Cancer Inform 2015;14(Suppl 2):139–45.

[32] Yang L, Cui Y, Shen J, et al. Antitumor activity of SA12, a novel peptide, on SKBr-3 breast cancer cells via the mitochondrial apoptosis pathway. Drug Des Devel Ther 2015;9:1319–30.

[33] Viola K, Kopf S, Huttary N, et al. Bay11-7082 inhibits the disintegration of the lymphendothelial barrier triggered by MCF-7 breast cancer spheroids; the role of ICAM-1 and adhesion. Br J Cancer 2013;108:564–9.

[34] Gonzalez-Bengtsson A, Asadi A, Gao H, et al. Estrogen enhances the expression of the polyunsaturated fatty acid Elongase Elovl2 via ERalpha in breast cancer cells. PLoS One 2016;11:e0164241.

[35] Durso DF, Bacalini MG, Sala C, et al. Acceleration of leukocytes' epigenetic age as an early tumor and sex-specific marker of breast and colorectal cancer. Oncotarget 2017;8:23237–45.

[36] Shajahan-Haq AN, Cook KL, Schwartz-Roberts JL, et al. MYC regulates the unfolded protein response and glucose and glutamine uptake in endocrine resistant breast cancer. Mol Cancer 2014; 13:239.

[37] Notte A, Rebucci M, Fransolet M, et al. Taxol-induced unfolded protein response activation in breast cancer cells exposed to hypoxia: ATF4 activation regulates autophagy and inhibits apoptosis. Int J Biochem Cell Biol 2015;62:1–4.

[38] Davis NM, Sokolosky M, Stadelman K, et al. Deregulation of the EGFR/PI3K/PTEN/Akt/mTORC1 pathway in breast cancer: possibilities for therapeutic intervention. Oncotarget 2014;5:4603–50.

[39] Guichard SM, Curwen J, Bihani T, et al. AZD2014, an inhibitor of mTORC1 and mTORC2, is highly effective in ER+ breast cancer when administered using intermittent or continuous schedules. Mol Cancer Ther 2015;14:2508–18.

[40] Lu W, Li Y. Salinomycin suppresses LRP6 expression and inhibits both Wnt/beta-catenin and mTORC1 signaling in breast and prostate cancer cells. J Cell Biochem 2014;115:1799–807.

[41] Rennhack J, Andrechek E. Conserved E2F mediated metastasis in mouse models of breast cancer and HER2 positive patients. Oncoscience 2015;2:867–71.

[42] Kawamoto H, Koizumi H, Uchikoshi T. Expression of the G2-M checkpoint regulators cyclin B1 and cdc2 in nonmalignant and malignant human breast lesions: immunocytochemical and quantitative image analyses. Am J Pathol 1997;150:15–23.