Identification and Validation of a Five-gene Signature Associated with Overall Survival in Breast Cancer Patients

Xiaolong Wang
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Chen Li
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Tong Chen
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Hanwen Zhang
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Ying Liu
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Dianwen Han
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Yaming Li
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Zheng Li
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Dan Luo
Department of Pathology Tissue Bank, Qilu Hospital of Shandong University, Jinan, Shandong, China
Ning Zhang
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China
Bing Chen
Department of Pathology Tissue Bank, Qilu Hospital of Shandong University, Jinan, Shandong, China
Lijuan Wang
Department of Pathology Tissue Bank, Qilu Hospital of Shandong University, Jinan, Shandong, China
Wenjing Zhao
Department of Pathology Tissue Bank, Qilu Hospital of Shandong University, Jinan, Shandong, China
Qifeng Yang (✉ qifengy_sdu@163.com)
Department of Breast Surgery, Qilu Hospital of Shandong University, Jinan, Shandong, China.
Department of Pathology Tissue Bank, Qilu Hospital of Shandong University, Jinan, Shandong, China.
Research Institute of Breast Cancer, Shandong University, Jinan,

Research

Keywords: Breast cancer, Bioinformatics, LASSO Cox, Prognostic biomarkers, Riskscore, Individualized therapy

Posted Date: September 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74127/v1

Abstract

Background

Recent years, attributed to early detection and new therapies, the mortality rates of breast cancer (BC) decreased. Nevertheless, the global prevalence was still high and the underlying molecular mechanisms were remained largely unknown. The investigation of prognosis-related genes as the novel biomarkers for diagnosis and individual treatment had become an urgent demand for clinical practice.

Methods

Gene expression profiles and clinical information of breast cancer patients were downloaded from The Cancer Genome Atlas (TCGA) database and randomly divided into training (n = 514) and internal validation (n = 562) cohort by using a random number table. The differentially expressed genes (DEGs) were estimated by Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. In the training set, the gene signature was constructed by the least absolute shrinkage and selection operator (LASSO) method based on DEGs screened by R packages. The results were further tested in the internal validation cohort and the entire cohort. Moreover, functions of five genes were explored by MTT, Colony-Formation, scratch and transwell assays. Western blot analysis was used to explore the mechanisms.

Results

In the training cohort, a total of 2805 protein coding DEGs were acquired through comparing breast cancer tissues (n = 514) with normal tissues (n = 113). A risk score formula involving five novel prognostic associated biomarkers (EDN2, CLEC3B, SV2C, WT1 and MUC2) were then constructed by LASSO. The prognostic value of the risk model was further confirmed in the internal validation set and the entire set. To explore the biological functions of the selected genes, *in vitro* assays were performed, indicating that these novel biomarkers could markedly influence breast cancer progression.

Conclusion

We established a predictive five-gene signature, which could be helpful for prognosis assessment and personalized management in breast cancer patients.

Introduction

Breast cancer is the most common diagnosed cancer and the leading cause of cancer-related death in women across the world [1]. According to the Cancer Statistics 2020, around 276,480 cases of female breast cancer were diagnosed in US with the expectation of 42,170 deaths [2]. Due to the early detection

and the progression in diagnosis and treatments, the mortality rate of breast cancer had declined over the past decades [3]. However, for the patients who progressed to metastasis or chemoresistance, the prognosis were still poor [4, 5]. Thus, there was an urgent need for the construction of a reliable risk model to evaluate the prognosis of breast cancer patients and identification of novel therapeutic targets for individual treatment.

Dysregulation of genes played crucial roles in various biological processes [6]. For the limitation on statistical property of single biomarkers, it was indicated by various studies that multigene signatures provided by systematic analysis could act as more accurate predictive biomarkers than the conventional clinicopathologic characteristics for the risk stratification [7, 8]. 21-gene signature was developed to evaluate the risk of distant and local recurrence, and estimate the benefit of chemotherapy for the ER-positive breast cancer [9]. 70-gene signature has been proved to improve the prediction of clinical prognosis for the early-stage breast cancer patients [10]. A two-gene epigenetic signature was shown to accurately predict the response of triple-negative breast cancer patients to the neoadjuvant chemotherapy [11]. Therefore, identification of novel multigene signatures played a critical role to ameliorate the prognosis of breast cancer patients and provide better treatment strategies for the high-risk population.

Over the past decades, in-depth gene sequencing and bioinformatics provided us the chance to identify novel diagnostic parameters and guide the individual treatment optimization for various illnesses [12–14]. Gene expression microarray was an effective method to show large-scale data at genomic levels, and rapid progression of bioinformatics make it possible to mine more reliable biomarkers [15]. The Cancer Genome Atlas (TCGA) was an open, public, large-scale database, which containing abundant raw data for cancer researches [16]. In our present study, based on the mRNA expression profiles acquired from the TCGA databases, a prognosis-associated gene signature was constructed by the LASSO Cox regression model [17, 18]. Also, the selected biomarkers were proved to play central roles in breast cancer progression by *in vitro* assays. EDN2, CLEC3B, SV2C, WT1 and MUC2 could serve as potential biomarkers affecting the prognosis in breast malignancy and provide us a new angle for better understand of molecular network in breast cancer progression [19].

Materials And Methods

Data source

The mRNA expression profile of breast cancer patients used to identify DEGs were derived from TCGA (http://tcga-data.nci.nih.gov/tcga/) on October 1, 2018，which contained 113 normal breast tissues and 1109 breast tumor tissues. 1076 patients with clinical information were enrolled in this study. TCGA databases was open-access and publicly available. The present study followed the data access policy and publishing guidelines.

The selection of differentially expressed genes

To identify the genes differentially expressed in the breast cancer tissues and normal tissues, the raw data of mRNA expression was normalized. Gene counts were converted into TPM (transcripts per million mapped reads) values and log2-transformed. R package “limma” was then used to screen the differentially expressed genes (DEGs). The screening conditions for differentially expressed genes were using the following criteria: |fold change (FC)| > 3 and adjusted false-discovery rate (FDR) < 0.05 was applied to find the upregulated and downregulated mRNAs. Adjust P value < 0.05. R package “pheatmap” was used to drawn the heatmap.

Functional Analysis

The Gene Ontology (GO) analysis and the Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis were widely used methods for systematic assessment of biological functional studies on high-throughput genomics data [20-22]. In this study, functional enrichment analyses of the GO analysis and KEGG analysis were performed by FunRich, an open access, standalone tool for functional enrichment and network analysis [23]. The molecular function, cellular component, biological process and biological pathway of DEGs were estimated. The “ggplot2” package for R software was used to analyze the data.

Construction of gene-related risk model for breast cancer

The least absolute shrinkage and selection operator (LASSO) method was a commonly used method for regression with high-dimensional predictors [24]. In this study, Lasso was used to obtain the most strongly survival-associated genes in the training set. The R packages “survival,” and “glmnet” were applied to perform a lasso regression analysis. The mRNA-related gene signature was expressed as follows:

risk score = (coefficient $_{gene\ 1}$ × status of gene 1) + (coefficient $_{gene\ 2}$ × status of gene 2) + … + (coefficient $_{gene\ n}$ gene n× status of gene n) [25].

Survival analysis

We analyzed the overall survival of patients by the Kaplan-Meier method. The R package “survival” and “survminer” were applied to construct the Kaplan–Meier survival plots (the difference in survival rates among different groups was measured and p < 0.05 was considered significant in the survival analysis).

Cell Culture

The human breast cancer cell lines MDA-MB-231 and MDA-MB-468 used in this study were purchased from American Type Culture Collection (ATCC, Manassas, VA, USA), and routinely maintained in DMEM/high glucose medium (Gibco-BRL, Rockville, IN, USA) with 10% fetal bovine serum (Haoyang Biological Manufacture, Tianjin, China), and 1% penicillin-streptomycin at a 37℃ cell culture incubator with 5% CO_2.

Quantitative Real-Time PCR (qPCR) and transfection

TRIzol reagent (Invitrogen, Carlsbad, CA, USA) was used to extract RNA from breast cancer cells. PrimeScript reverse transcriptase reagent kit (TaKaRa, Shiga, Japan) was used to reverse-transcribe mRNAs into cDNAs. The specific primers were as follows: EDN2 forward: 5'-CGTCCTCATCTCATGCCCAAG-3', EDN2 reverse: 5'- AGGCCGTAAGGAGCTGTCT-3', CLEC3B forward: 5-'CCCAGACGAAGACCTTCCAC-3', CLEC3B reverse: 5'-CGCAGGTACTCATACAGGGC-3', SV2C forward: 5'-TCCTACAGTCGGTTCCAAGAT-3', SV2C reverse: 5'-GGCCTCACCATTATAGGTTTCTC-3', WT1 forward: 5'-CACAGCACAGGGTACGAGAG-3', WT1 reverse: 5'-CAAGAGTCGGGGCTACTCCA-3', MUC2 forward: 5'-AGGATGACACCATCTACCTCAC-3', MUC2 reverse: 5'-CATCGCTCTTCTCAATGAGCA-3', Transfection was conducted with Lipofectamine 2000 (Invitrogen). The overexpression plasmids were purchased from Vigene Biosciences (Shandong, China).

MTT (3-(4,5-dimethyl-2-thiazolyl)-2,5–diphenyl-2H-tetrazolium Bromide) Assay

Cell proliferation assay was determined using MTT (Sigma, St. Louis, MO, USA) according to the instructions. MDA-MB-231 and MDA-MB-468 cells were plated into 96-well cell culture plates with at least three replicate wells for each group. Afterwards, 20 μL of MTT (5 mg/mL in PBS) was added to each well and incubated for another 6 h at 37°C. The supernatants were then aspirated carefully and 100 μL of dimethyl sulfoxide (DMSO) was added to each well. Absorbance values were measured using a Microplate Reader (Bio-Rad, Hercules, CA, USA) at 490 nm.

Colony-Formation Assay

EDN2, CLEC3B, SV2C and WT1 overexpression cells and control cells were digested by trypsin and seeded in a 6 cm dish at a density of 1000 cells/dish. MDA-MB-231 cells were cultured for 15 days, MDA-MB-468 cells were cultured for 30 days. Then the clones were washed by PBS, fixed with methanol for 5 min, stained with 0.1% crystal violet. Three independent experiments were performed for the same conditions.

Scratch assay

After transfected with selected genes, MDA-MB-231 and MDA-MB-468 cells were seeded on a 24-well plate at the density of 1×10^5 /well and 1.5×10^5/well, respectively. A straight-line cell-free “scratch” was created by pipette tips and a horizontal line at the back of the plate was drawn as reference point to guarantee the same area of image acquisition. After washed with PBS to remove debris, the plate was incubated in 5% CO2 at 37°C. The migration speed was measured by calculating the difference in the distances between the two edges of the scratch.

Transwell Assay

Cell migration ability was evaluated by transwell assay using Transwell chamber with pore size of 8.0μm (Millipore) according to the manufacturer’s instructions. 1×10^5 MDA-MB-231 cells and 1.5×10^5 MDA-MB-468 cells were suspended in serum-free medium and plated on upper wells. The medium containing 20%

FBS was added to the lower chamber as chemoattractant. MDA-MB-231 cells were culture for 12h, and MDA-MB-468 were cultured for 30h. After being fixed with methanol for 5 min, the chambers were stained with 1% crystal violet solution for 5 min. Then, the cells in the lower chamber were observed under an inverted microscope. Three independent experiments were performed for the same conditions.

Western Blotting

The MDA-MB-231 and MDA-MB-468 cells were transfected with EDN2, CLEC3B, SV2C, WT1, and the control cells were transfected with pENTER plasmid. Subsequently, after washing with ice-cold PBS, the proteins of distinctively treated cells were collected and lysed in lysis buffer in the presence of protease inhibitors. After centrifugation at 12,000 rpm for 20 min at 4˚C, the supernatant was collected. 30 μg protein were separated by 10% SDS-PAGE and transferred (100 V, 2 h) onto polyvinylidene fluoride (PVDF) membranes (Millipore, Bedford, MA, USA). After blocking with 5% nonfat milk for 1 h, the membranes were incubated overnight at 4°C with the primary antibodies. After washing with TBS-T, the membrane was labeled with the secondary antibody and protein spots were visualized by ECL. β-actin was used as the endogenous control.

Statistical analysis

All the experiments were conducted for the same conditions in triplicate. Statistical analyses in the study were performed with SPSS (version 23.0) and GraphPad Prism 8.0. Kaplan-Meier plots was used to conduct survival analysis. Significant differences were evaluated by student's t-test and one-way analysis of variance (ANOVA). P value < 0.05 were considered statistically significant.

Results

Clinicopathological features of breast cancer patients

1076 breast cancer patients with clinical information were collected from the TCGA database. By using a random number table, samples were divided into training (n = 514) and internal validation (n = 562) cohort. The detailed demographic and clinicopathological characteristics of the patients involved in the training, internal and entire datasets were shown in Table 1. The percentages of patients at clinical stages I, II, III and IV in the training cohort were 16.7%, 55.3%, 23.3% and 1.6%, respectively. The data was 16.4%, 58.2%, 22.6% and 2.1% in the internal validation cohort. Among them, 354 patients in the training cohort and 389 patients in the internal validation cohort had prognosis information. The follow-up periods encompassed the different pathological stages of breast cancer. The median follow-up periods were 592 days (range, 2-6434 days) and 631 days (range, 1-7125 days), respectively, in the training and internal validation sets. In the training cohort, the median age was 58 years (range, 27-90 years). And, in the internal validation cohort, the median age was 58 years (range, 26-90 years). During the follow-up, in training cohort, 48/354 (13.6%) patients died, whereas in internal validation cohort, 45/389 (11.6%) patients died.

Identification of DEGs

The exploration process of this study is shown in Figure 1. Firstly, 60483 genes were initially screened in the training cohort. Thresholds were set as fold change>3 and FDR<0.05 to investigate the differentially expressed genes between normal group and tumor group (Figure 2A). Volcano plots were used to show the differentially expressed genes (Figure 2B). 4805 genes had differential expressions between normal and tumor tissues, which consisted of 1269 upregulated genes and 3536 downregulated genes. 2805 DEGs were with protein coding functions.

Enrichment analyses of DEGs

To further understand the function of DEGs, the differentially expressed mRNAs were incorporated into functional annotation analyses. The molecular processes during the progression of breast cancer were investigated through GO enrichment analyses and KEGG pathway analyses.

The upregulated mRNAs associated with molecular function were enriched in the modulation of structural constituent of chromatin, chemokine activity and metallopeptidase activity (Figure 3A). In terms of cellular component, the upregulated genes were tightly corresponding to chromosome passenger complex, Ndc80 complex and condensed chromosome kinetochore (Figure 3B). Additionally, in the analysis on biological process, spindle assembly, chromosome segregation and negative regulation of enzyme activity were the most enriched terms mediated by upregulated genes (Figure 3C). From the prospective of biological pathways, the high expression genes were closely related to aurora B signaling, mitotic prometaphase and PLK1 signaling events (Figure 3D)

Meanwhile, downregulated genes related to molecular function were enriched in lipase activity, serotonin degradation and chemokine activity (Figure 4A). Through the investigation on cellular component, the most enriched terms were voltage-gated sodium channel complex, lipid particle and keratin filament (Figure 4B). In the exploration of biological process, downregulation genes in breast cancer were enriched in regulation of membrane potential, lipid storage and regulation of transport (Figure 4C). Besides, analysis on biological pathways proved that downregulated genes were associated with noradrenaline and adrenaline degradation, serotonin degradation and HSL-mediated triacylglycerol hydrolysis (Figure 4D).

Construction and validation of risk prognostic scoring system in the training set.

2805 protein coding genes were further selected using LASSO regression analysis, and cross validation was used to select the penalty parameters (Figure 5A). Five genes were identified for the construction of the prognostic model by a multivariate Cox regression analysis (Figure. 5B). The genes obtained in the above steps were inserted into the multivariate Cox regression model. The expression values of five independent prognostic factors and their correlation coefficients in a multivariate regression model were then used to construct prognostic signatures. Detailed information and the significance of survival prediction by the five genes are presented in Table 2.

Risk score= (expression status of EDN2× 0.014) C (expression status of CLEC3B× -0.196) + (expression status of SV2C× 0.227) + (expression status of WT1× 0.075) + (expression status of MUC2× 0.113).

Of five biomarkers, three genes (EDN2, WT1 and MUC2) were upregulated in breast cancer samples while CLEC3B and SV2C were decreased (Figure 5C-G). Moreover, the protein expression of the five genes were further explored in the Human Protein Atlas (HPA) and the representative pictures of them were shown in Figure 5H-I.

In the training cohort, the distributions of risk score of breast cancer patients and the relationships between risk score and survival time were visualized in Figure 6A and 6B. The patients' mRNA expression levels of selected genes were shown in Figure 6C. Patients in the training cohort were then assigned to a high- or low-risk score group using the cut-off value (0.09) obtained with the "survival" and "survminer" package. 198 (56%) patients in the training cohort were categorized to the high-risk group (RS > 0.09) and 156 (44%) to the low-risk group (RS ≤ 0.09). High-risk patients also had markedly shorter OS (HR 1.88, 95% CI 1.07–3.31, $p < 0.05$) vs. low-risk patients in the training cohort (Figure 6D).

Validation of the five genes-model in the internal testing group and entire group.

To assess the stability and reliability of the five genes signature, the result was also tested in the internal validation cohort and entire cohort. In the internal validation cohort, according to the same median risk score that acquired from the training group, 389 breast cancer patients with follow-up information were divided into high- and low-risk group. Figure 7A and 7B showed the distributions of risk score of breast cancer patients and the relationships between risk score and survival time. Expressions of the five genes in the risk score formula in the testing group were provided in the Figure 7C. As for the survival analysis, using the "survival" and "survminer" package, the optimal cut-off value (0.14) was obtained to divide the patients in the internal validation cohort into high-risk group (160) and low-risk group (229). The relationship between the distribution of risk score and clinical information indicated that the higher patients ranking, predicted poorer overall survival (HR 1.78, 95% CI 0.96-3.31, $p < 0.05$) (Figure 7D).

In Figure 8A and 8B, patients in the entire group (n = 743) were divided into the high-risk group and the low-risk group in the same way by the median risk score. The risk score distributions and the survival status were exhibited. The expressions of five selected genes in 743 patients were also shown in Figure 8C. Calculated by the "survival" and "survminer" package, the optimal cut-off value was 0.09. Thus, patients in the entire cohort were then categorized to high-risk group (n=407) and low-risk group (n=336). The results of survival analysis were showed in the Kaplan-Meier plot (Figure 8D). With the extension of survival time, the survival rate of the high-risk group became lower, and had a poor prognosis effect (HR 1.72, 95% CI 1.15-2.59, $p < 0.01$).

Gain-of-function assay of selected genes

To evaluated the influence of these genes on breast cancer progression, MDA-MB-231 and MDA-MB-468 cell lines were used to assess the biological roles of the selected genes. EDN2, CLEC3B, SV2C and WT1

were overexpressed by transfection to conduct gain-of-function assay. The overexpression efficiency of selected genes was verified by qPCR (Figure 9A, B). We tested cell proliferative viability using the MTT assay in MDA-MB-231 and MDA-MB-468 cells. As shown in Figure 9C and 9D, overexpression of CLEC3B and WT1 could significantly promote the growth in both cell lines, while EDN2 and SV2C had no obvious influence on cell proliferation. The results were then further validated by the clone formation assay (Figure 9E, F). Overexpression of CLEC3B and WT1 could dramatically promote the formation of colonies in breast cancer cells.

Furthermore, cell migration assays were used to evaluate the regulation effects of selected genes on cell migration. As evidenced by scratch assay and transwell assay, the mobility of MDA-MB-231 and MDA-MB-468 cells overexpressed CLEC3B and WT1 were considerably increased compared with control group (Figure 9G-J). On the contrary, transfected with SV2C could significantly inhibited cell migration in both breast cancer cell lines, while the function of EDN2 was slight.

Influence of selected genes on epithelial-mesenchymal transition (EMT) signaling pathway in breast cancer

The EMT represented a biological process during which polarized epithelial cells lost cell identity and experienced various biochemical alterations that allowed it to assume mesenchymal phenotypes [26]. Normally observed during embryonic development, EMT could also be involved in various pathological conditions. Once hijacked by cancer cells, EMT often led to enhanced migration capability, acquisition of resistance to apoptosis, and increased cell proliferation [26-29]. Thus, we examined the role of selected genes in EMT signaling pathway in breast cancer cells. As shown in Figure. 9, gain-of-function of EDN2, CLEC3B and WT1 markedly increased the ZEB1 and β-catenin in MDA-MB-231. Besides, CLEC3B and WT1 could also enhance the expression of snail (Figure 10A). By contrast, SV2C seemed to play a key role as tumor suppressor in EMT signaling pathway. MDA-MB-231 cells transfected with SV2C showed low expression of EMT markers, such as ZEB1, vimentin, β-catenin and snail. In MDA-MB-468, EDN2, CLEC3B and WT1 were proved to be able to upregulate the protein level of β-catenin (Figure 10B). CLEC3B and WT1 transfection led to higher expression level of N-Cadherin. Moreover, a markedly increase of snail was also observed in the WT1 overexpressed MDA-MB-468 cells.

Discussion

As one of the most malignant tumors in women, breast cancer was a heterogeneous disease with diverse subtypes. Each subtype had distant biological and clinical characteristics [30]. It was of great importance to investigate the underlying molecular pathogenesis of breast cancer and find reliable prognostic biomarkers for the identification of patients with high risk [31]. Microarray data had been proved as an effective tool in identification of gene biomarkers, which was a crucial step for tumor assessment [32]. In the present study, gene expression profiles of breast cancer samples and corresponding normal tissue were download from TCGA together with clinical information. Since the absence of suitable independent cohorts for validation, patients with breast cancer were randomly divided into training and internal

validation cohorts to ensure the stability of the prognostic model. Candidate genes were prescreened by analysis of the differentially expressed genes between breast cancer and control samples.

In total, 2805 DEGs were identified. To further investigate the molecular mechanisms involved in breast cancer, GO and KEGG analysis were performed [33, 34]. As shown in our data, the upregulated DEGs were mainly enriched in the DNA repair machinery, enhanced cell mobility and limitless replicative potential. PLK1 was a serine/threonine protein kinase, which played a critical role in the regulation of cell cycle and chemoresistance [35, 36]. Our data demonstrated that upregulated DEGs were highly associated with PLK1 signaling pathway. This indicated that the dysregulation of PLK1 signaling pathway contributed to the prognosis of breast cancer patients. As for the downregulated DEGs, the enriched terms included correspond to immune response, chemokine activity and regulation of metabolism. These findings were also consistent with previous breast cancer studies [37–39].

Penalized methods had aroused much attention as a novel predicting tool for high accuracy and good feasibility [40]. L_1-penalty, also known as LASSO, was the most widely used penalty in high-dimensional cancer classification [41]. Recent studies had showed that LASSO could be used as an effective tool in exploration of potential biomarkers in breast cancer. A 6-KIFs-based risk score (KIF10, KIF15, KIF18A, KIF18B, KIF20A, KIF4A) reported by Li et al.[42] was proved to be associated with the prognosis in patients with breast cancer. Immune related index in breast cancer were also found through LASSO by Xie et al.[43] and Zheng et al.[44]. These researchers indicate that gene signatures could serve as risk factors for cancer management and played a vital role in predicting cancer prognosis.

In our study, to further explore the prognosis related biomarkers in breast cancer, LASSO Cox regression model was performed. We screened out five protein coding genes (EDN2, CLEC3B, SV2C, WT1 and MUC2) significantly corresponding to the overall survival time of patients with breast cancer in the training group. Compared to single biomarker alone, the risk score consisted of the coefficient and expression status of multiple genes markedly increased the reliability and accuracy of diagnosis result. Thus, a 5-genes signature was established as potential biological indicators for breast cancer diagnosis and prognosis. The gene signature was also tested in the internal validation cohort and the entire cohort. The Kaplan-Meier plot showed the significant difference of overall survival between the high- and low-risk group. The 5-gene prognostic model was expected to work as an auxiliary predicting tool in the individual management of breast cancer.

Through the literature search, it was found that several biomarkers related to the gene signature were reported to be involved in the process of cancer development and progression. EDN2 had been reported to be an oncogene overexpressed in various malignancies, which corelated to cell differentiation, proliferation, migration and resistance to chemotherapy [45–49]. However, functions of EDN2 in breast cancer had not been reported. Besides, CLEC3B seemed to play distinct roles in different human cancers. While functioning as tumor suppresser in lung cancer [50], expression of CLEC3B was proved to be related to poor prognosis in colorectal cancer and gastric cancer [51, 52]. WT1 was firstly identified as a tumor suppressor gene in nephroblastoma [53]. However, it was demonstrated by subsequent studies that

WT1 was related to the disruption of EMT signaling pathway and docetaxel resistance in breast cancer, high expression of WT1 was also corresponding to lower overall survival [53, 54]. Functioned as an oncogene, MUC2 was highly expressed in mucin secreting breast cancers and played a pivotal role in regulating cell proliferation, metastasis and apoptosis [55].

To further validated the functions of the biomarkers, *in vitro* assays were then performed to evaluated the influence of the selected genes on proliferation and metastasis. We proved that EDN2 could be associated with the protein level in EMT signaling pathway. It was found that CLEC3B and WT1 could markedly enhance the capability of growth and migration in breast cancer cell lines. Meanwhile, overexpressing SV2C could decrease the cell mobility *in vitro*. Detection of the protein levels further proved that changed migration ability was probably caused by the alteration of EMT signaling pathway. In the present study, we established a novel five-gene signature which was a promising tool in predicting breast cancer prognosis. Three of the genes in the five-gene signature were reported to be related to breast cancer for the first time. These potential biomarkers could be helpful for future investigation.

However, there were some limitations which need to be mentioned in this study. First, no external validation cohort was involved in this study and our data were all acquired from the TCGA database. Second, only overall survival was taken into consideration, and the quality of life of breast cancer patients were not covered. Besides, for the present research was retrospective, the predicting model should also be testified in the large-scale prospective studies. Thus, the results demanded to be further verified before applied into clinical practice.

Conclusion

In summary, a five-gene (EDN2, CLEC3B, SV2C, WT1, MUC2) based prognostic model was constructed and validated in the study, which was proved to be an accurate classifier for risk stratification and clinical decision-making. These selected genes were tightly corresponding to the progression of breast cancer and might serve as potential targets for future individual treatment.

Abbreviations

TCGA, The Cancer Genome Atlas; LASSO, the least absolute shrinkage and selection operator; DEGs, the differentially expressed genes; BC, breast cancer; GO, Gene Ontology; KEGG, the Kyoto Encyclopedia of Genes and Genomes; qPCR, Quantitative Real-Time PCR; MTT, 3-(4,5-dimethyl-2-thiazolyl)-2,5–diphenyl-2H-tetrazolium Bromide; OS, overall survival; RS, risk score; HR, hazard ratio.

Declarations

Availability of data and materials

The data of this study are from TCGA database.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Consent to publish was obtained from all authors.

Competing interests

The authors declare no competing interests.

Funding

Refer to Acknowledgement section.

Author Contributions

The authors contributed as follows: Conceptualization, Q.Y and X.W.; Data curation, C.L., T.C., H.Z., Y.L. and D.H.; Investigation, C.L., D.L. and N.Z.; Methodology, X.W., B.C., L.W. and W.Z.; Original Draft Preparation, C.L.; Review & Editing, X.W.; Figure Preparation and Editing, C.L. Y.L., Z.L., and X.W.; Supervision, Q.Y.; Project Administration, Q.Y.; Funding Acquisition, Q.Y. All authors reviewed the article and approved the final version of the manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 81272903; No. 81672613; No. 81874119), Key Research and Development Program of Shandong Province (No. 2016GSF201119), Shandong Science and Technology Development Plan (2016CYJS01A02), Special Support Plan for National High Level Talents (“Ten Thousand Talents Program”), Shandong Provincial Natural Science Foundation, China (No. ZR2019LZL003), and Clinical New Technology Developing Fund (No. 2018-7).

References

1. Sharma R. Breast cancer incidence, mortality and mortality-to-incidence ratio (MIR) are associated with human development, 1990–2016: evidence from Global Burden of Disease Study 2016. Breast Cancer. 2019;26(4):428–45.
2. Siegel R, Miller K, Jemal A, *Cancer statistics*, 2020. CA: a cancer journal for clinicians, 2020. **70**(1): p. 7–30.

3. Kroenke CH, et al. Postdiagnosis social networks breast cancer mortality in the After Breast Cancer Pooling Project. 2017;123(7):1228–37.
4. Chang E, et al. Association between prolonged metastatic free interval and recurrent metastatic breast cancer survival: findings from the SEER database. Breast cancer research treatment. 2019;173(1):209–16.
5. Weigel M, Dowsett M. Current and emerging biomarkers in breast cancer: prognosis and prediction. Endocrine-related Cancer. 2010;17(4):R245-62.
6. Li W, et al., *An Integrated Model Based on a Six-Gene Signature Predicts Overall Survival in Patients With Hepatocellular Carcinoma.* Frontiers in Genetics, 2019. **10**.
7. Buyse M, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006;98(17):1183–92.
8. Reis-Filho J, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet. 2011;378(9805):1812–23.
9. Sparano J, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. Journal of clinical oncology. 2008;26(5):721–8.
10. Cardoso F, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375(8):717–29.
11. Pineda B, et al. A two-gene epigenetic signature for the prediction of response to neoadjuvant chemotherapy in triple-negative breast cancer patients. Clinical epigenetics. 2019;11(1):33.
12. Latha NR, et al. Gene expression signatures: A tool for analysis of breast cancer prognosis and therapy. Crit Rev Oncol Hematol. 2020;151:102964.
13. Peng PL, et al. Identification of a novel gene pairs signature in the prognosis of gastric cancer. Cancer Med. 2018;7(2):344–50.
14. Bao X, et al. A novel epigenetic signature for overall survival prediction in patients with breast cancer. J Transl Med. 2019;17(1):380.
15. Zhang X, et al. Integrative transcriptome data mining for identification of core lncRNAs in breast cancer. PeerJ. 2019;7:e7821.
16. Network CGAR. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543–50.
17. Cui Q, et al., *A prognostic eight-gene expression signature for patients with breast cancer receiving adjuvant chemotherapy*. J Cell Biochem, 2019.
18. Xie X, et al. Identification of a 4-mRNA metastasis-related prognostic signature for patients with breast cancer. J Cell Mol Med. 2019;23(2):1439–47.
19. Wang H, Wu L, Wang H. *Development and verification of a personalized immune prognostic feature in breast cancer*. Exp Biol Med (Maywood), 2020: p. 1535370220936964.
20. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic acids research. 2015;43(Database issue):D1049–56.

21. Dutkowski J, et al. A gene ontology inferred from molecular networks. Nature biotechnology. 2013;31(1):38–45.
22. Kanehisa M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-d361.
23. Pathan M, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics. 2015;15(15):2597–601.
24. R, T., *Regression shrinkage and selection via the lasso*. Vol. Vol. 58. 1996: Journal of the Royal Statistical Society. Series B. 22.
25. Zhang JX, et al. Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. Lancet Oncol. 2013;14(13):1295–306.
26. Georgakopoulos-Soares I, et al. EMT Factors and Metabolic Pathways in Cancer. Front Oncol. 2020;10:499.
27. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009;119(6):1420–8.
28. Kalluri R, Neilson EG. Epithelial-mesenchymal transition and its implications for fibrosis. J Clin Invest. 2003;112(12):1776–84.
29. Wang Z, et al. DNER promotes epithelial-mesenchymal transition and prevents chemosensitivity through the Wnt/β-catenin pathway in breast cancer. Cell Death Dis. 2020;11(8):642.
30. Blows F, et al., *Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies.* PLoS medicine, 2010. **7**(5): p. e1000279.
31. Tian Z, et al., *An immune-related prognostic signature for predicting breast cancer recurrence*. Cancer medicine, 2020.
32. Wang Y, Li X. and R.J.I.t.o.c. Ruiz. Weighted General Group Lasso for Gene Selection in Cancer Classification. 2019;49(8):2860–73.
33. Kanehisa M, et al. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-62.
34. *Gene Ontology Consortium: going forward*. Nucleic Acids Res, 2015. 43(Database issue): p. D1049-56.
35. Barr F, Silljé H, Nigg E. Polo-like kinases and the orchestration of cell division. Nature reviews. 2004;5(6):429–40.
36. Giordano A, et al. Polo-like kinase 1 (Plk1) inhibition synergizes with taxanes in triple negative breast cancer. PloS one. 2019;14(11):e0224420.
37. Sledge GW Jr, Miller KD. Exploiting the hallmarks of cancer: the future conquest of breast cancer. Eur J Cancer. 2003;39(12):1668–75.
38. Kalimutho M, et al. Patterns of Genomic Instability in Breast Cancer. Trends Pharmacol Sci. 2019;40(3):198–211.

39. Duijf PHG, et al. Mechanisms of Genomic Instability in Breast Cancer. Trends Mol Med. 2019;25(7):595–611.
40. Algamal ZY, M.H.J.E.S.w A, Lee. *Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification*. 2015. 42(23): p. 9326–9332.
41. Tibshirani R, *Regression Shrinkage and Selection via the Lasso*. Vol. 58. 1996: Journal of Royal Statistical Society. Series B.
42. Li TF, et al. Overexpression of kinesin superfamily members as prognostic biomarkers of breast cancer. Cancer Cell Int. 2020;20:123.
43. Xie P, et al. Development of an Immune-Related Prognostic Signature in Breast Cancer. Front Genet. 2019;10:1390.
44. Zheng S, et al., *Identification and validation of a combined hypoxia and immune index for triple-negative breast cancer*. Mol Oncol, 2020.
45. Berry P, Burchill S, *Endothelins may modulate invasion and proliferation of Ewing's sarcoma and neuroblastoma.* Clin Sci (Lond), 2002. **103 Suppl 48**: p. 322 s-326 s.
46. Grimshaw MJ. Endothelins in breast tumour cell invasion. Cancer Lett. 2005;222(2):129–38.
47. Bagnato A, Spinella F, Rosanò L. Emerging role of the endothelin axis in ovarian tumor progression. Endocr Relat Cancer. 2005;12(4):761–72.
48. Wiesmann F, et al. Frequent loss of endothelin-3 (EDN3) expression due to epigenetic inactivation in human breast cancer. Breast Cancer Res. 2009;11(3):R34.
49. Yang H, et al. Paclitaxel Sensitivity of Ovarian Cancer Can be Enhanced by Knocking Down Pairs of Kinases that Regulate MAP4 Phosphorylation and Microtubule Stability. Clin Cancer Res. 2018;24(20):5072–84.
50. Sun J, et al. CLEC3B as a potential diagnostic and prognostic biomarker in lung cancer and association with the immune microenvironment. Cancer Cell Int. 2020;20:106.
51. Chen H, et al. High Intratumoral Expression of Tetranectin Associates with Poor Prognosis of Patients with Gastric Cancer after Gastrectomy. J Cancer. 2017;8(17):3623–30.
52. Zhu HF, et al. Cancer-associated fibroblasts promote colorectal cancer progression by secreting CLEC3B. Cancer Biol Ther. 2019;20(7):967–78.
53. Zhang Y, et al. The role of WT1 in breast cancer: clinical implications, biological effects and molecular mechanism. Int J Biol Sci. 2020;16(8):1474–80.
54. Miyoshi Y, et al. High expression of Wilms' tumor suppressor gene predicts poor prognosis in breast cancer patients. Clin Cancer Res. 2002;8(5):1167–71.
55. Astashchanka A, Shroka TM, Jacobsen BM. Mucin 2 (MUC2) modulates the aggressiveness of breast cancer. Breast Cancer Res Treat. 2019;173(2):289–99.

Tables

Table 1

Clinical characteristics of patients with breast cancer involved in this study			
	Training set (n=514)	Internal validation set (n=562)	Entire set (n=1076)
Age			
≥65	171 (33.3)	163 (29.0)	334 (31.0)
65	343 (66.7)	399 (71.0)	742 (69.0)
Sex			
Male	509 (99.0)	555 (98.8)	1064 (98.9)
Female	5 (1.0)	7 (1.2)	12 (11.1)
Primary tumor location			
Left-sided	268 (52.1)	292 (52.0)	560 (52.0)
Right-sided	246 (47.9)	269 (47.9)	515 (47.9)
Unexamined	0 (0)	1 (0.1)	1 (0.1)
Clinical risk group			
Stage I	86 (16.7)	92 (16.4)	178 (16.5)
Stage II	284 (55.3))	327 (58.2)	611 (56.8)
Stage III	120 (23.3)	127 (22.6)	247 (22.9)
Stage IV	8 (1.6)	12 (2.1)	20 (1.9)
Unexamined	16 (3.1)	4 (0.7)	20 (1.9)
T stage			
T1	137 (26.6)	138 (24.6)	275 (25.6)
T2	289 (56.2)	335 (59.6)	624 (58.0)
T3	62 (12.1)	71 (12.6)	133 (12.4)
T4	23 (4.5)	17 (3.0)	40 (3.7)
Unexamined	3 (0.6)	1 (0.2)	4 (0.3)
N stage			
N0	243 (47.3)	260 (46.3)	503 (46.7)
N1	167 (32.5)	190 (33.8)	357 (33.2)
N2	64 (12.5)	56 (10.0)	120 (11.1)

N3	28 (54.5)	47 (8.4)	75 (7.0)
Unexamined	12 (2.3)	9 (1.5)	21 (2.0)
M stage			
M0	423 (82.3)	475 (84.5)	898 (83.5)
M1	10 (1.9)	12 (2.1)	22 (2.0)
Unexamined	81 (15.8)	75 (13.4)	156 (14.5)
ER status			
ER positive	377 (73.3)	415 (73.8)	792 (73.6)
ER negative	113 (22.0)	120 (21.4)	233 (21.7)
Unexamined	24 (4.7)	27 (4.8)	51 (4.7)
PR status			
PR positive	317 (61.7)	367 (65.3)	684 (63.6)
PR negative	170 (33.1)	168 (29.9)	338 (31.4)
Unexamined	27 (5.2)	27 (4.8)	54 (5.0)
Her-2 status			
Her-2 positive	92 (17.9)	100 (17.8)	192 (17.8)
Her-2 negative	344 (66.9)	395 (70.3)	739 (68.7)
Unexamined	78 (15.2)	67 (11.9)	145 (13.5)
Margin status			
Margin positive	32 (6.2)	46 (8.2)	78 (7.2)
Margin negative	434 (84.4)	470 (83.6)	904 (84.0)
Close	13 (2.5)	13 (2.3)	26 (2.4)
Unexamined	35 (6.9)	33 (5.9)	68 (6.4)

Data are n (%).

Table 2

Gene symbol	Full name	Coefficient
EDN2[1]	Endothelin 2	0.014
CLEC3B[2]	C-type lectin domain family 3 member B	-0.196
SV2C	Synaptic vesicle glycoprotein 2C	0.227
WT1[3]	WT1 transcription factor	0.075
MUC2[4]	Mucin 2, oligomeric mucus/gel-forming	0.113

[1]Also known as: ET-2, ET2, PPET2. [2]Also known as: TN, TNA. [3]Also known as: AWT1, GUD, NPHS4, WAGR, WIT-2, WT33. [4]Also known as: MLP, MUC-2, SMUC.

Figures

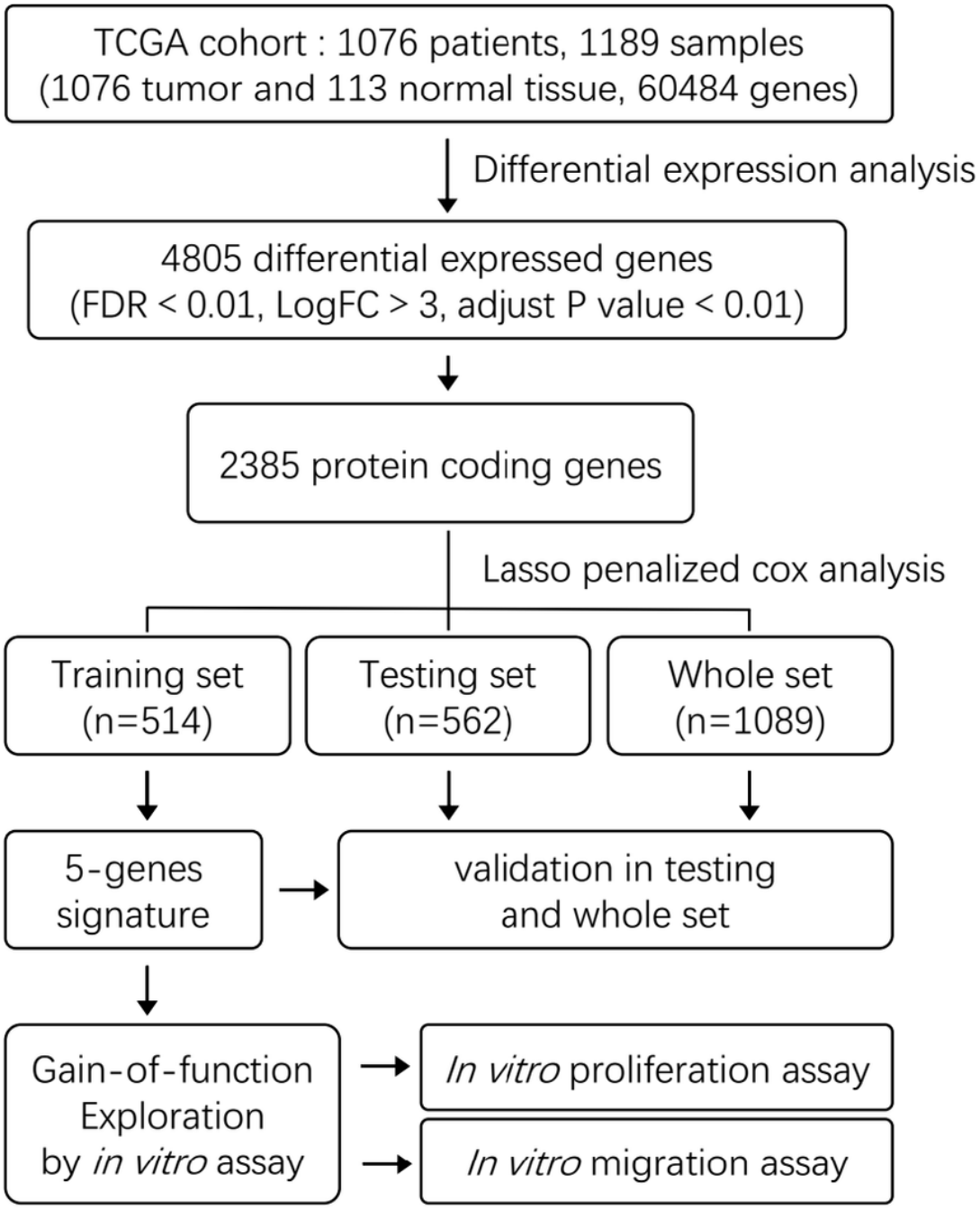

Figure 1

The flow chart showing the scheme of the study on five-gene prognostic signatures for breast cancer.

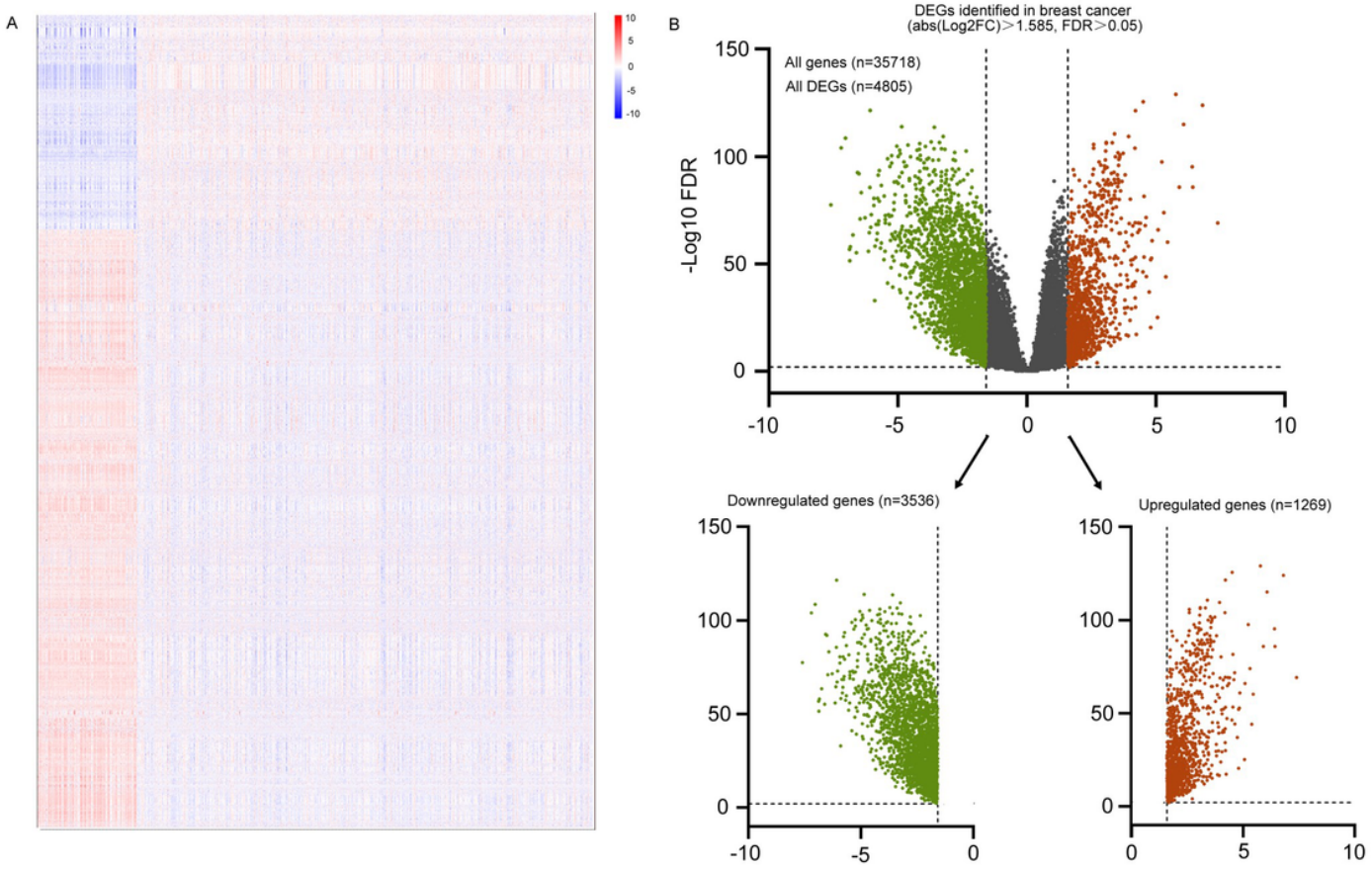

Figure 2

Heatmap and volcano plot were used to show the DEGs in breast cancer. (A) Heatmap represented mRNAs differentially expressed between breast cancer and normal breast tissues based on microarray analysis. (B) Volcano plot represented all differential expressed genes, green indicated downregulated genes, and red indicated all upregulated genes.

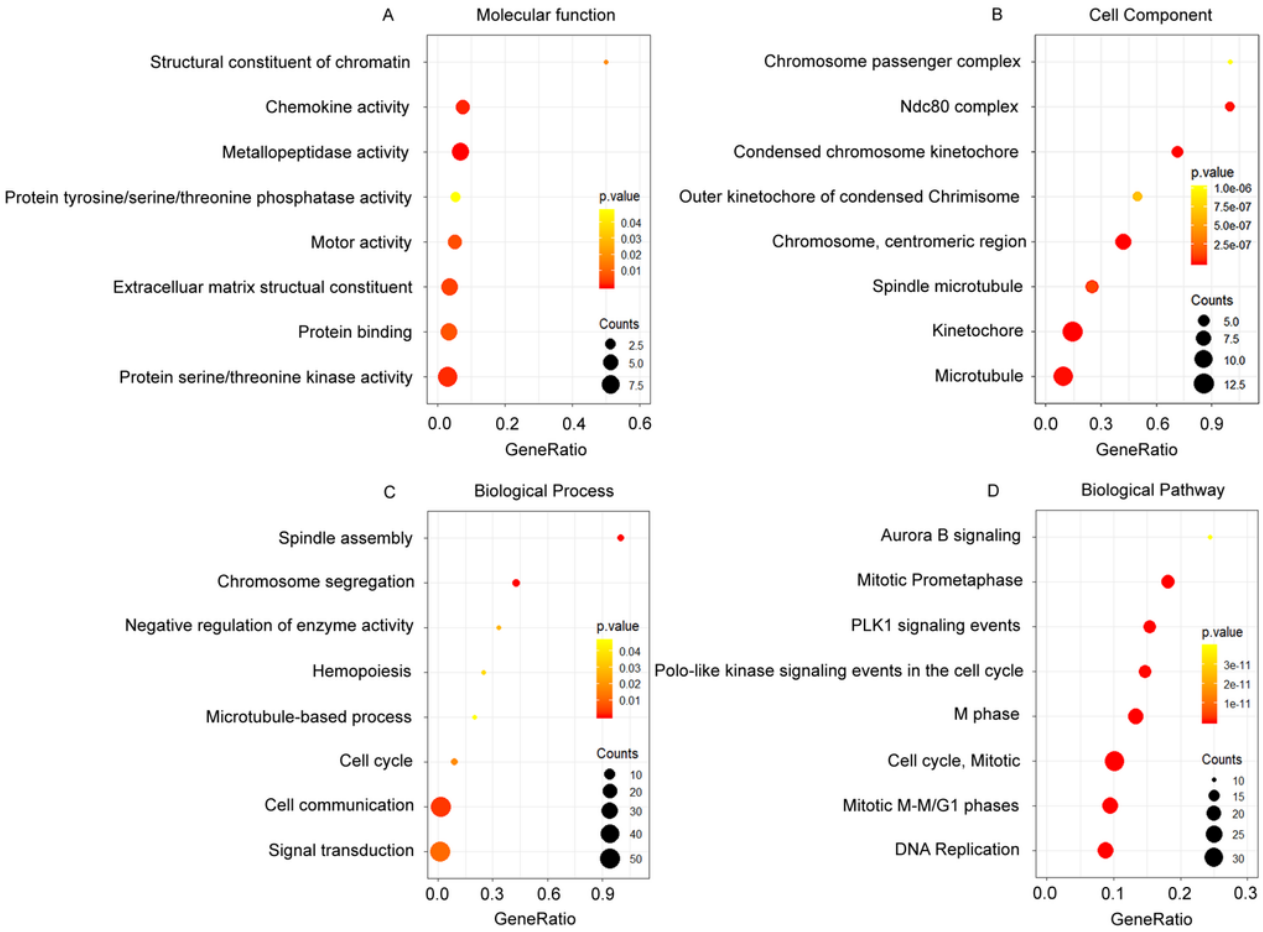

Figure 3

Functional enrichment analysis of the upregulated genes. (A) Enrichment of molecular function. (B) Enrichment of cellular component. (C) Enrichment of biological process. (D) Enrichment of Kyoto Encyclopedia of Genes and Genomes.

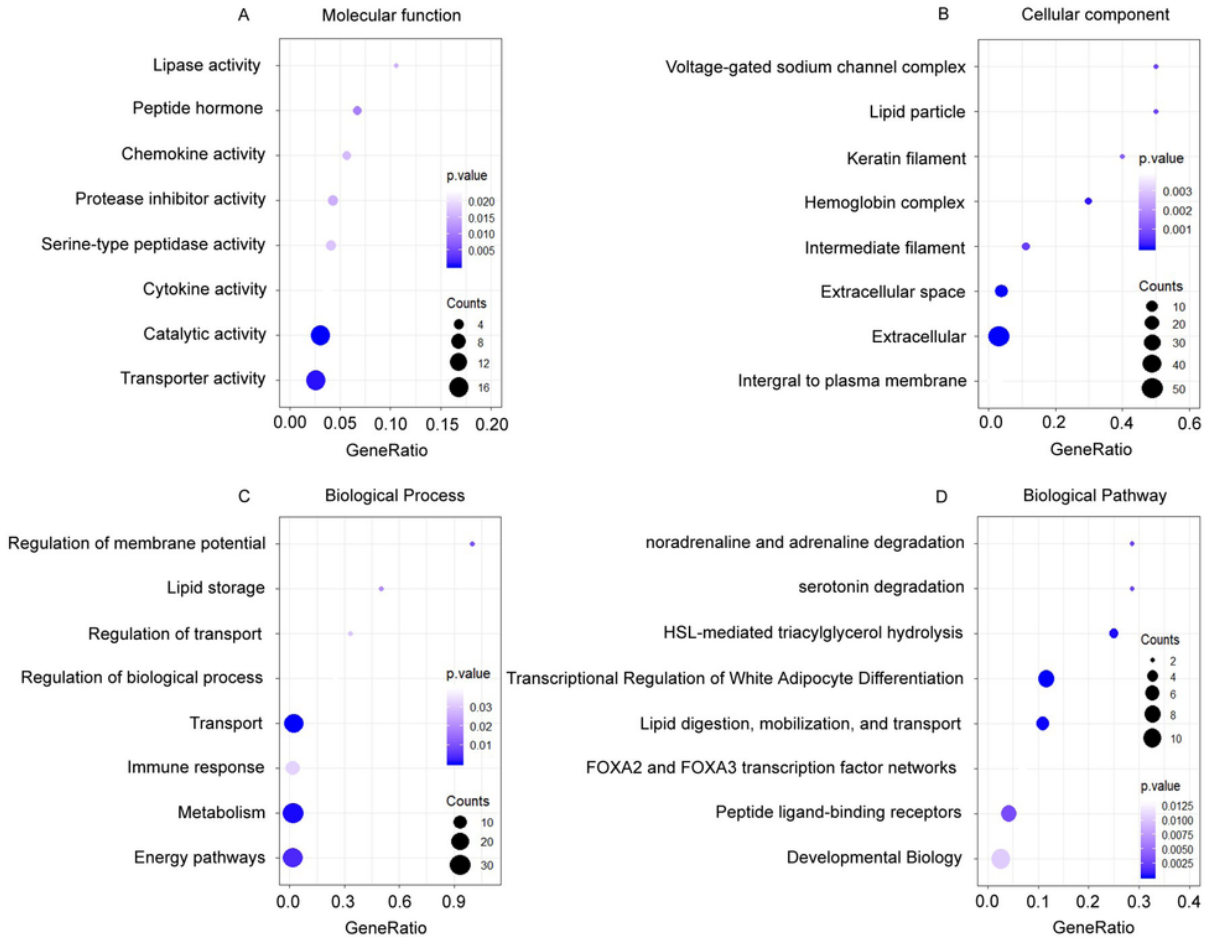

Figure 4

Functional enrichment analysis of the downregulated genes. (A) Enrichment of molecular function. (B) Enrichment of cellular component. (C) Enrichment of biological process. (D) Enrichment of Kyoto Encyclopedia of Genes and Genomes.

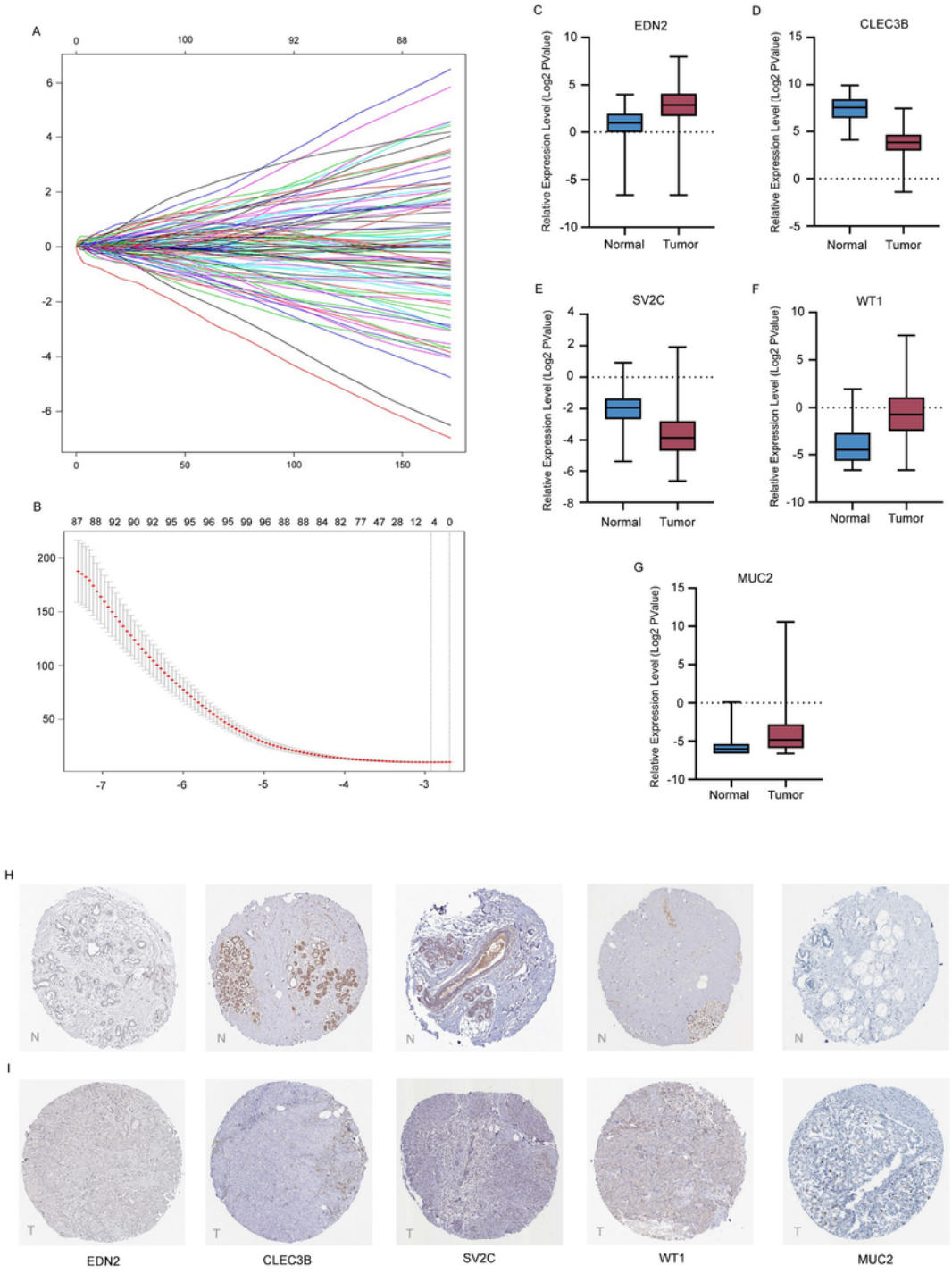

Figure 5

Construction of the five-gene prognostic model and validation of expression of the five genes in breast cancer. (A) LASSO Cox regression model. (B) Cross validation of LASSO regression. (C-G) The mRNA expression levels of selected genes in the training cohort. (H-I) The representative protein expression of the five genes in breast cancer tumor tissue and normal tissue. Data was obtained from the human protein atlas.

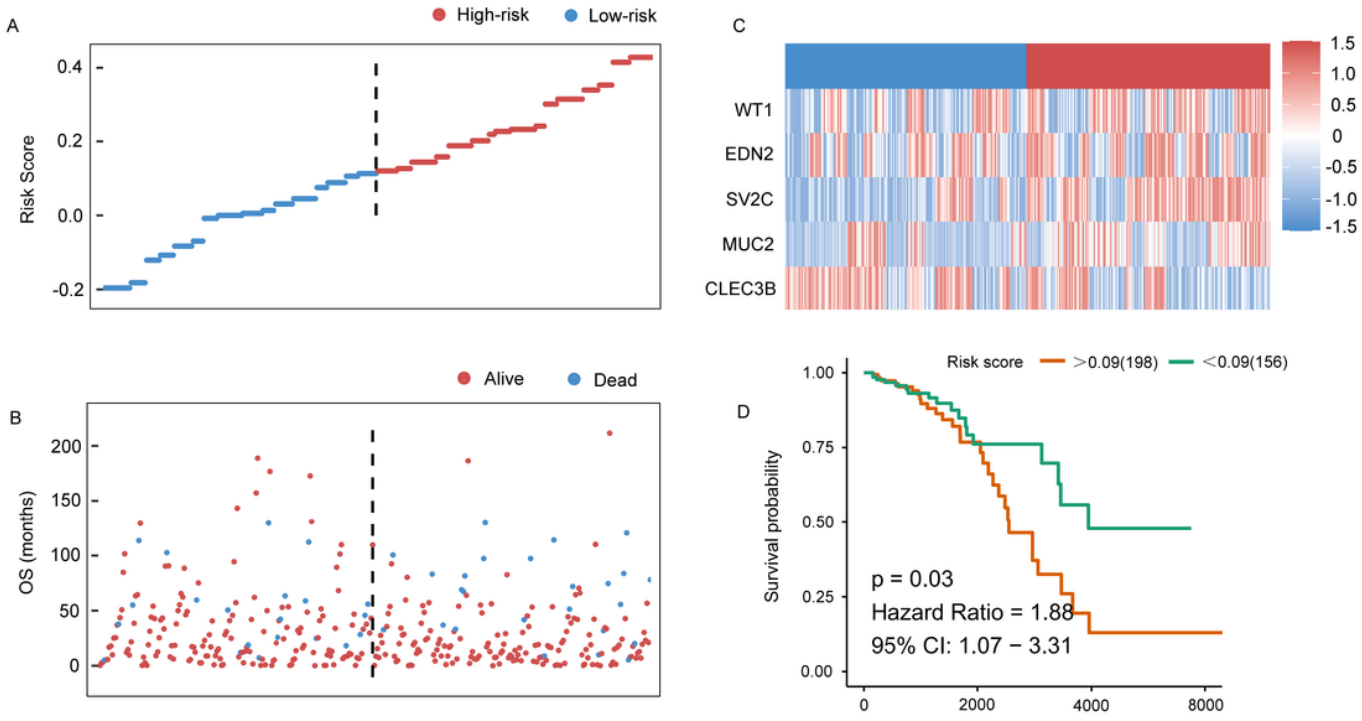

Figure 6

Evaluating the predictive power of five-gene signature in the training group. (A) Distribution of risk score. (B) Survival status of breast cancer patients in the training group. (C) Heatmap of the prognosis-associated gene expression profiles in the training cohort. (D) Kaplan-Meier plot of the high- and low-risk groups in the training group.

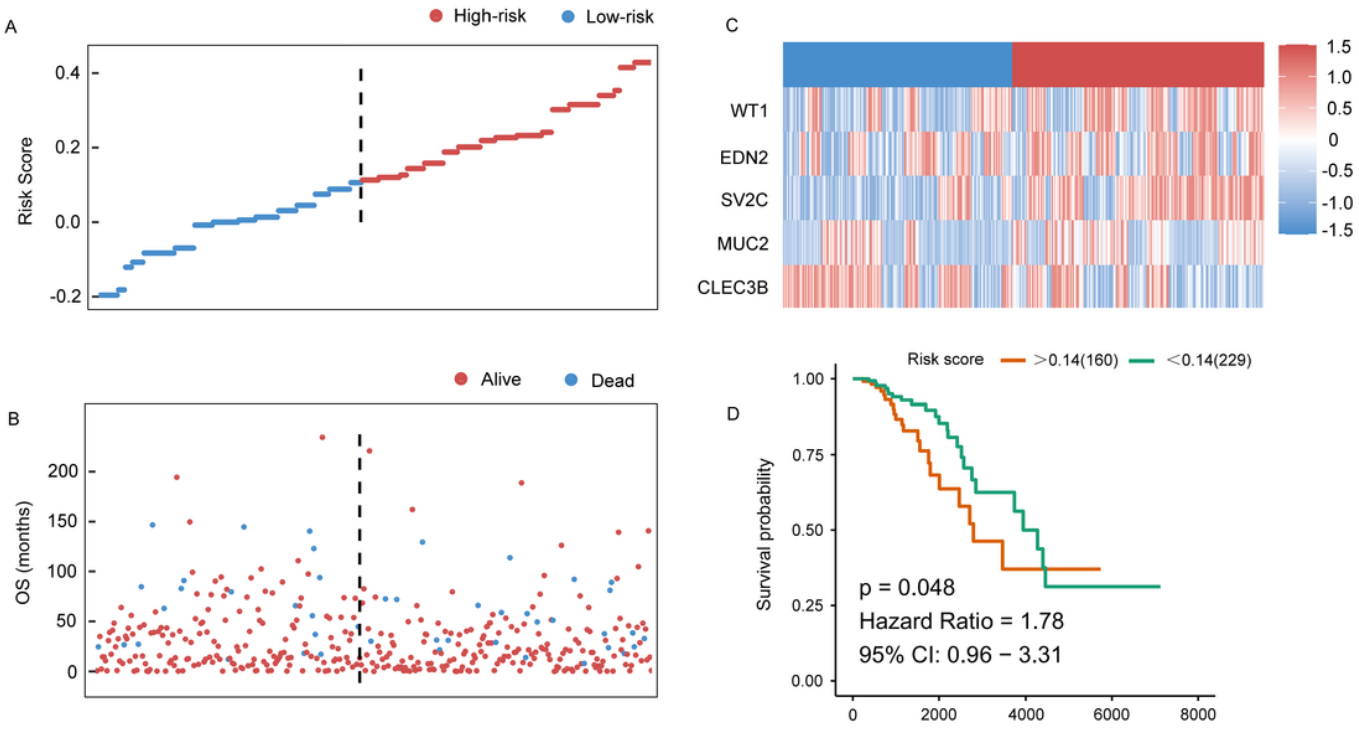

Figure 7

Evaluating the predictive power of five-gene signature in the internal validation group. (A) Distribution of risk score. (B) Survival status of breast cancer patients in the internal validation group. (C) Heatmap of the prognosis-associated gene expression profiles in the internal validation cohort. (D) Kaplan-Meier plot of the high- and low-risk groups in the internal validation group.

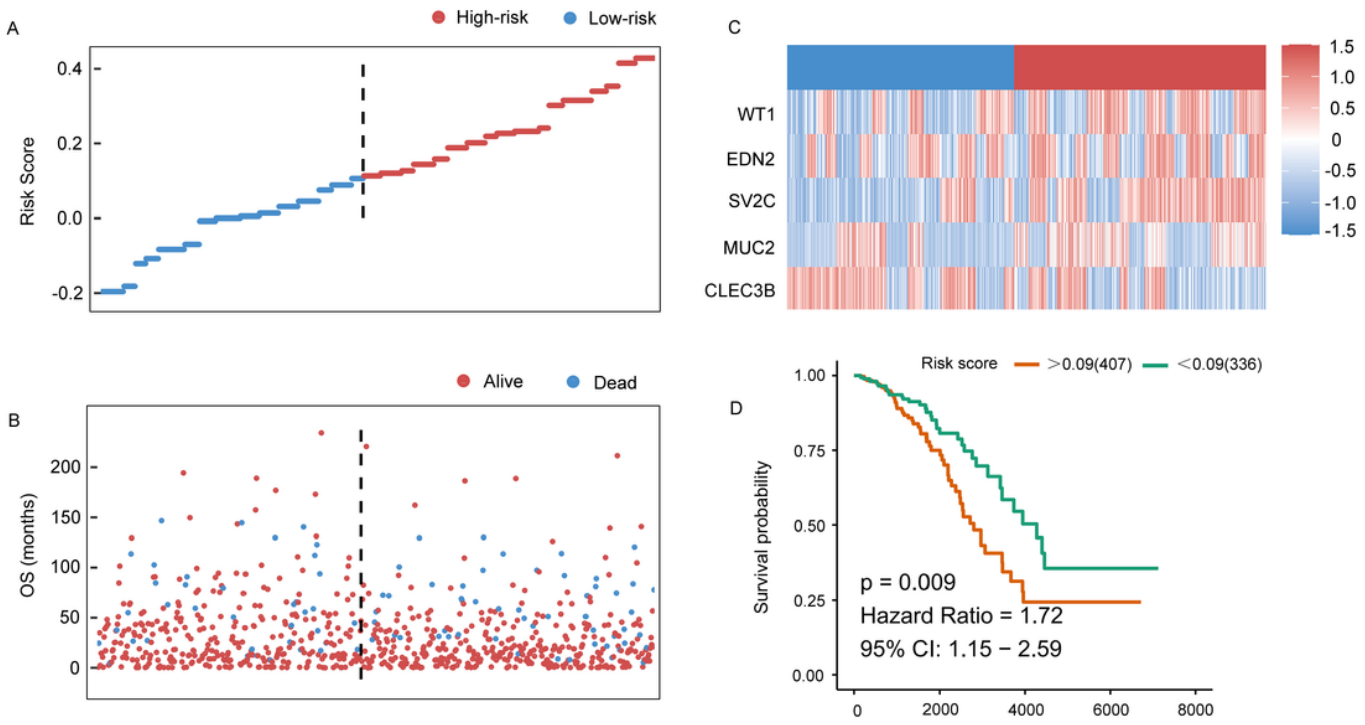

Figure 8

Evaluating the predictive power of five-gene signature in the entire group. (A) Distribution of risk score. (B) Survival status of breast cancer patients in the entire group. (C) Heatmap of the prognosis-associated gene expression profiles in the entire cohort. (D) Kaplan-Meier plot of the high- and low-risk groups in the entire group.

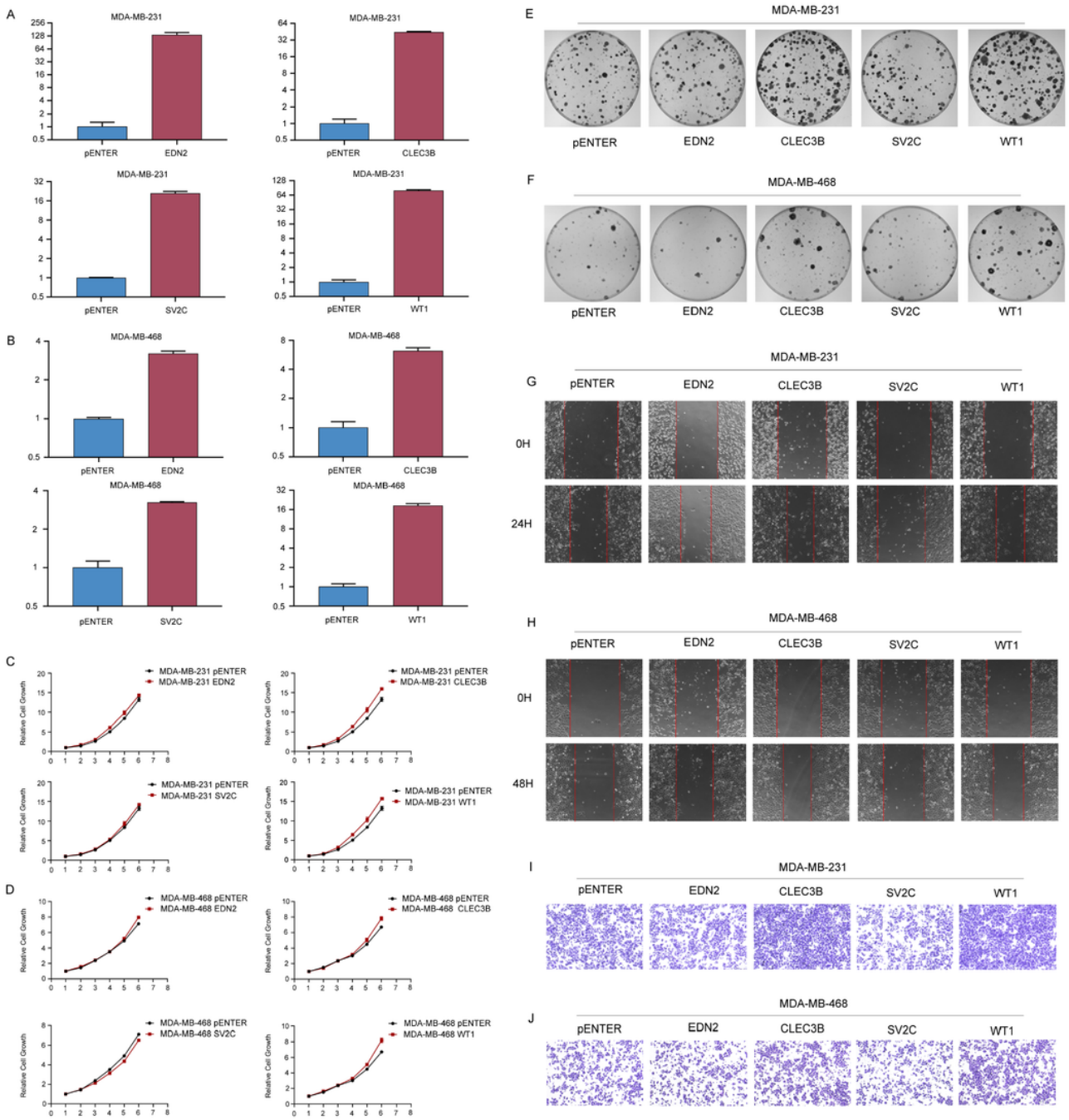

Figure 9

Gain-of-function assay of selected genes regulating cell proliferation and metastasis. (A-B) The overexpression efficiency of selected genes in MDA-MB-231 and MDA-MB-468 cells. (C-D) Effect of selected genes on cell proliferation was tested by MTT in MDA-MB-231 and MDA-MB-468 cells. (E-F) Effect of selected genes on cell proliferation was tested by colony formation assay in MDA-MB-231 and MDA-MB-468 cells. (G-H) Effect of selected genes on cell migration was tested by scratch assay in MDA-

MB-231 and MDA-MB-468 cells. (I-J) Effect of selected genes on cell migration was tested by transwell assay in MDA-MB-231 and MDA-MB-468 cells.

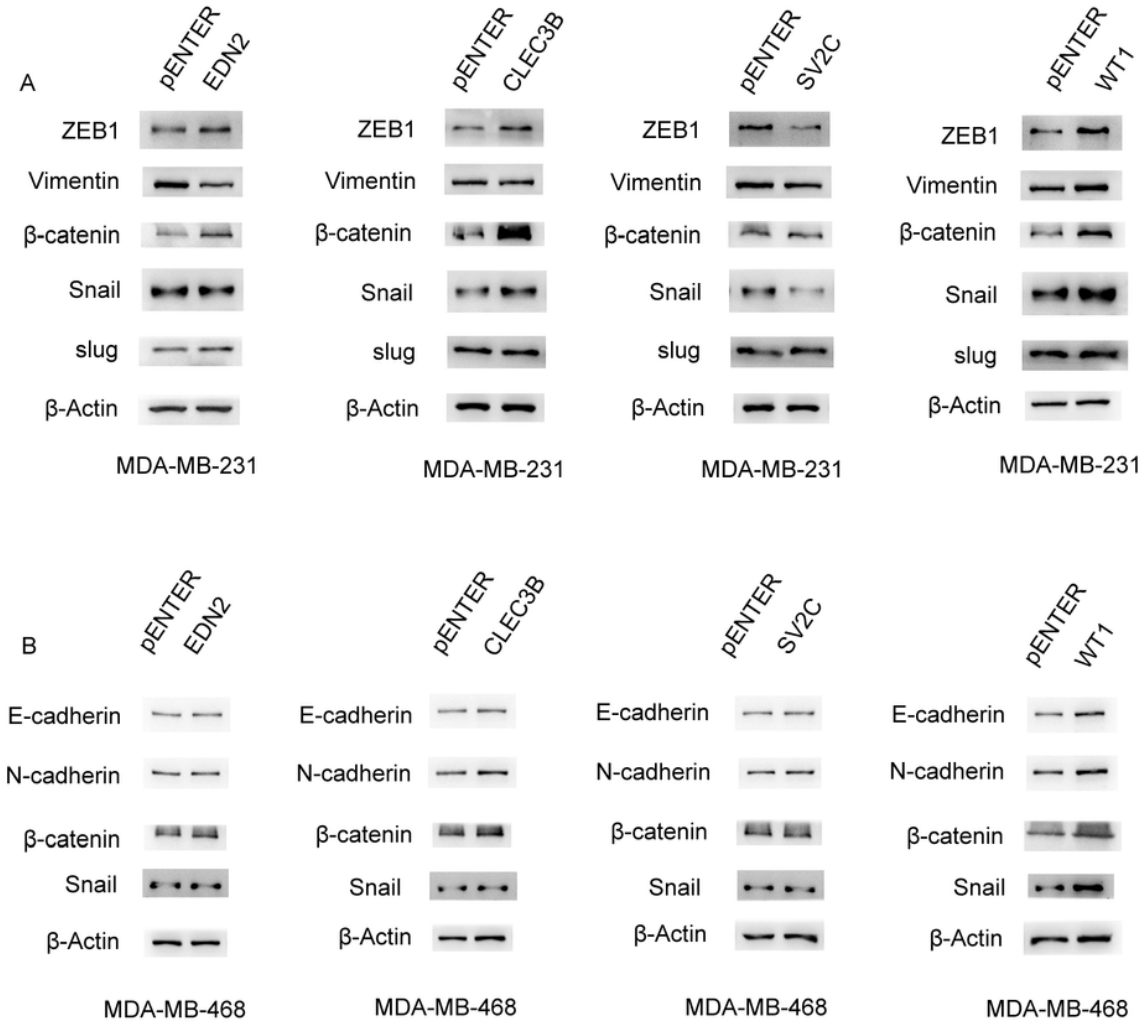

Figure 10

Influence of selected genes on EMT signaling pathway in breast cancer. (A) Effect of selected genes on protein level of EMT signaling pathway was measured by Westernblot assay in MDA-MB-231. (B) Effect of selected genes on protein level of EMT signaling pathway was measured by Westernblot assay in MDA-MB-468.