Development a Prognostic Model Integrating lncRNA/ mRNA novel Biomarkers Identified by Bioinformatics Analysis and Experiments in Breast Cancer

Jinrong Wei
Second Affiliated Hospital of Soochow University https://orcid.org/0000-0002-7009-2073
Qianshu Dou
Second Affiliated Hospital of Soochow University
Futing Ba
Second Affiliated Hospital of Soochow University
Guo-Qin Jiang (jiang_guoqin@suda.edu.cn)
Second Affiliated Hospital of Soochow University https://orcid.org/0000-0002-5294-9360

Primary research

Keywords: breast cancer, prognosis, lncRNA, TME, CST1, C20orf85, ceRNA

Posted Date: March 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-251621/v1

Abstract

Purpose: The purpose of this study is to established a prognosis model based on the expression profiles of lncRNAs and mRNAs for breast cancers.

Methods: Single Variable Cox Proportional Risk Regression analysis and difference analysis were applied to screen survival-related and differently expressed lncRNAs and mRNAs between tumor and normal tissues from TCGA data. GO and KEGG analysis were applied for top 30 survival-related genes. LncRNA/mRNA co-expressed network was constructed based on correlation analysis. LASSO analysis and Multivariate Stepwise Cox Regression analysis were applied to establish the prognosis model. RT-PCR experiments were applied to verify the correctness of the analysis results. Relative components of the TME in breast cancers with high and low risk groups were analysed by **xCell** and Cox proportional risk regression analysis. The ceRNA network was constructed by calculating the Pearson correlation coefficient (PCC) for miRNA-mRNA and miRNA-lncRNA using paired miRNA, mRNA, and lncRNA expression profile data.

Results: Venn diagrams showed that there were 60 genes and 54 lncRNAs that were differently expressed and related with survival. Through lncRNA/mRNA co-expressed network construction, 19 lncRNA and 16 mRNA hub genes were gained. The genes were then included in LASSO and multivariate Cox proportional hazard regression analysis, and finally, 3 lncRNAs (LINC01497, LINC02766, LINC02528) and 2 mRNAs (C20orf85, CST1) were selected as prognosis predictive genes. According to the median risk score of the 5 candidates, patients were divided into high-risk group and low-risk group. The results of RT-PCR were consistent with the analysis results. The proportions of Adipocytes, Endothelial cells, HSCs, Fibroblasts were significantly lower in low risk score tissues compared with the high risk score tissues, while the proportions of M1 macrophages, MSCs, Th2 cells were significantly higher. A lncRNA-miRNA-mRNA ceRNA network containing 3 lncRNAs, 2 mRNAs, and 158 miRNAs was finally constructed, preliminarily revealed a proper mechanism of the 5 molecules playing important roles in breast cancer progression and prognosis prediction.

Conclusion: We found that LINC01497, LINC02766, LINC02528 and C20orf85, CST1 may serve as a powerful prognostic tool to optimize the prognosis evaluation system of breast cancer.

Introduction

Breast cancer is the most common cancer and the most frequent cause of cancer death among women worldwide. The incidence is rising in most countries and is projected to rise further over the next 20 years despite current efforts to prevent and treatment of the disease[1, 2]. Breast cancer is a complex, heterogeneous disease classified into hormone-receptor-positive, human epidermal growth factor receptor-2 overexpressing (HER2+) and triple-negative breast cancer (TNBC) based on histological features[3]. However, the great heterogeneity of breast cancer makes it impossible to firmly predict the prognosis of each subtype patients according to the conventional prognostic factors currently employed.

Considering that most oncological treatments have short- and long-term toxic effects, new methods capable of offering a more precise prognosis need to be developed. Especially find out the subset of patients with early-stage breast cancer at high risk for recurrence who have the urgent need for additional therapeutic schemes[4]. However, such clinical and pathological criteria are of relative efficacy and have potential limitations in identifying the risk of disease relapse. Therefore, it is necessary to explore effective prognostic biomarkers to help optimize the prognosis evaluation system of breast cancer. In recent years, a new prognostic molecular classification has been developed based on the grouping of tumors according to their gene expression profiles. The individualisation of the diagnosis of patients with breast cancer based on molecular and gene expression studies is bringing about a veritable revolution in our understanding of the biology of the disease and will lead to the application of more specific treatments, thereby improving patient survival with lesser toxicity and increased economic savings.

Long non-coding RNA (lncRNA) genes are an important population of non-coding RNAs with key roles in tumorigenesis process. Evidences suggest that they can be classified as tumor suppressor genes or oncogenes according to their functions and expression pattern in tumoral tissues. Their physiological and pathological functions have been shown to be exerted via their interactions with microRNAs (miRNAs), mRNAs, proteins and genomic DNA[5]. Their important roles in the regulation of cancer-related pathways in addition to deregulation of their expression in a number of cancers have suggested that they can be used as markers for cancer detection and prognosis, as well as targets for cancer treatment[6]. Deregulation of a number of lncRNAs has been detected in breast cancer samples and cell lines[7]. In addition to miRNAs, lncRNAs have been shown to regulate the plasticity of CSCs and been suggested as possible targets for anti-CSC treatments, such as HOTAIR, XIST, MALAT and H19 [8, 9]. Besides, the role of lncRNAs in the epithelial-to-mesenchymal transition (EMT) programs has also been partly elucidated[10]. But the role of lncRNAs in predicting the outcome of breast cancer has not been elucidated. Currently, emerging bioinformatics resources are assisting these types of analyses. Large-scale public data with gene expression and clinical information, complete biological databases, and sophisticated high-throughput data analysis methods together provide opportunities for identifying broader prognostic features in breast cancer biology. Here, the expression patterns of lncRNAs in breast cancer, as well as their significance in prognosis are discussed.

In order to explore prognosis-related lncRNAs and genes in breast cancer, we performed a multiperspectives, multi-dimensional analysis of a large number of breast cancer samples using massive bioinformatics and machine learning methods, such as differential analysis, weighted gene co-expression network analysis, univariate COX analysis, LASSO analysis, functional analysis, Cox proportional risk regression analysis etc. We applied difference analysis and single factor Cox proportional risk regression analysis to screen differently expressed and related to prognosis lncRNAs and genes from 1104 tumoral and 113 normal breast tissues in The Atlas Cancer Genome (TCGA) breast cancer dataset. Then, the intersection genes gained by the collection analysis were applied to the Least Absolute Shrinkage and Selection Operator (LASSO) and the Multivariate stepwise Cox regression analysis to establish the prognosis model. Finally, five biomarkers (3 lncRNAs and 2 mRNAs) of breast cancer were identified that were thought to be probably important prognostic features in breast cancer. According to the median risk

score, patients were divided into high-risk group and low-risk group, and survival analysis was conducted to evaluate the prognostic value of risk score. To verify the accuracy of the above bioinformatics analysis, RT-PCR experiments were applied. Besides, we applied xCell analysis for TME components comparison between high-risk and low-risk groups. Finally, in order to initially clarify the mutual regulation between the five biomarkers, we construct a ceRNA network. Isolation and identification of differentially expressed lncRNAs and genes not only helped to discover the function of genes, but also helped to reveal the pathogenesis of the disease.

Methods

Data acquisition

RNA-sequencing data, updated clinical data, and sample information of breast invasive cancer (BRCA) cohort were downloaded from the TCGA data portal (https://portal.gdc. cancer.gov/); A total of 1,276 specimens, consisting of 1,104 cancer samples and 162 normal samples, were obtained from TCGA. Use R (version 3.6.0) software to standardize and process data. According to the gene encoding proteins, the genes are divided into coding genes and non-coding genes, and corresponding counts are extracted for subsequent difference analysis.

Differentially expressed genes (DEGs) screening

We used the R package (R-3.6.3 https://www.r-project.org/) of "DESeq2" to analyze the differences between mRNA and lncRNA from the normal tissue samples and breast cancer tissues in the expressing data. The significance analysis with false discovery rate (FDR) < 0.05 and |log2 fold change (FC)| ≥ 2 was applied to select lncRNAs and mRNAs for further analysis.

Survival-related genes generation

The association between lncRNAs and mRNAs and patients' overall survival was analyzed in the TCGA cohort. R packages “survival”and “survminer” were applied for the survival analysis. 1,104 tumor samples out of 1,276 cases had a detailed survival time record, with time span from0 to 23.6 years, which were used for survival analysis. A univariate Cox proportional regression model was used to calculate the association between the expression of each lncRNA, mRNA and OS. P value < 0.05 & HR1.5(HR > 1.5 or HR < 0.67) were used to analyze the effect of lncRNA or mRNA on the prognosis of breast cancer. R package “ggplot2” is used for the volcanic diagram. Kaplan–Meier method was used to plot the survival curve, and log rank as the statistical significance test; p < 0.05 was considered significant.

GO and KEGG enrichment analysis

To search the key genes and pathways that were associated with breast cancer survival, the DEGs (differently expressed genes) were uploaded to the Database for Annotation, Visualization and Integrated Discovery (DAVID) to screen enriched biological themes, particularly GO terms and KEGG pathways [28, 29]. P < 0.05 was set as the cut-off criterion. R package “Cluster Profiler” was used to carry out Gene Ontology (GO) enrichment analysis including biological process(BP), cell components(CC) and molecular functions(MF) for the differentially expressed survival-related genes. The same tool is also used for the enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis.

Venn analysis

The collection analysis software venny 2.1.0 (https://bioinfogp.cnb.csic.es/tools/venny/index.html) was used to draw the Venn diagram, the significantly prognosis-related lncRNAs, mRNAs (p value < 0.05 & HR1.5 (HR >1.5 or HR < 0.67)) and differentially expressed lncRNAs, mRNAs (FDR < 0.05 & p value < 0.05 & | log2FC | > 2) take the intersection.

Construction of the lncRNA/mRNA co-expression network

To construct the lncRNA/mRNA co-expression network, we calculated the Pearson correlation coefficient and R value to evaluate lncRNA-mRNA correlation. The network construction procedure includes: (1) Preprocess data: the same mRNAs with different transcripts taking the median value represent the gene expression values, without special treatment of lncRNAs expression value. (2) Screen data: remove the subset of data according to the lists showing the differential expression of lncRNAs and mRNAs. (3) Calculate the Pearson correlation coefficient and use R value to calculate the correlation coefficient between lncRNAs and mRNAs. (4) Screen by Pearson correlation coefficient: select the Pearson correlation coefficient ≥ 0.99 or ≤−0.99 as the meaningful value and draw (correlation>0.5) the lncRNA/mRNA co-expression network by using the cytoscape (3.5.1) software program.

Candidate genes selection

Hub genes, highly interconnected with nodes in a module, have been considered functionally significant. In our study, an interesting module was chosen, and hub genes were defined by module connectivity. We defined genes with the node connectivity > 2 (total nodes) as the hub nodes in lncRNA/mRNA co-expression network and were chosen as the candidates to be further analyzed and validated.

Feature Selection by LASSO analysis

LASSO is a machine learning algorithm in which both variable selection and regularization occur simultaneously. This model uses a penalty to shrink regression coefficients toward zero, a number of variables will be eliminated because their coefficients will shrink to exactly zero[11]. According to the

collection analysis results by software venny, a batch of genes in modules that were closely related to the prognosis of breast cancer was obtained. In the present study, the survival-related lncRNAs and mRNAs identified were included in the LASSO regression analysis by using the R package “glmnet”, and the penalty parameter “lambda” was selected to choose the best model based on leave-one-out cross-validation, which is more suitable than tenfold cross-validation for a smaller number of samples. Finally, we extracted variables with nonzero coefficients and their corresponding coefficients.

Multi Cox proportional regression analysis and stepwise Cox

Combined with the overall survival rate of breast cancer patients in TCGA, the R package “survival” was used to perform multivariate COX regression analysis and stepwise COX multivariate regression analysis on the lncRNAs and mRNAs selected by LASSO to obtain lncRNAs and mRNAs (p value < 0.05). To make our model more optimized and practical, Muti Cox proportional regression analysis and a stepwise Cox proportional hazards regression model was used. Finally, a risk score formula was calculated by taking into account of the expression of optimized lncRNAs and mRNAs and correlation estimated Cox regression coefficients: Risk Score = 0.43203*LINC01497+0.69806*LINC02766-0.54019*LINC02528-0.25365*C20orf85-0.078*CST1. Patients with breast cancer were classified into the high- or low-risk group by ranking the given risk score. Differences in overall survival (OS) between the high- and low-risk group were assessed using Kaplan–Meier method and two-tailed log-rank test. A Cox proportional hazards regression model was used to identify independent prognostic factors. P < 0.05 was set as significant difference in all statistical methods.

Quantitation of the 5 biomarkers

Total RNA was extracted from the frozen tissues of patients using TRIzol reagent (Invitrogen, USA) following the manufacturer's protocol. Purified total RNA was reverse transcribed using the PrimeScript RT reagent kit (Takara, Japan). Quantitative real-time PCR (qPCR) of the 3 lncRNAs was performed using the Power SYBR Green PCR Master Mix (lnRcute lncRNA qPCR detection kit) according to the manufacturer's guidelines. The primers for the 3 lncRNA primers were as follows: LINC01497:forward: AAATCAAGGTGTTGGCTGGGCTAC, reverse: GTGTTGCTGGCTCCGAAGATGG; LINC02766: forward: AGGAAGTAGGCGGTGTGGAGTG, reverse: GACAGAGTGGGCGGGAGGAG; LINC02528, forward: ACCTACCGAGAGACCTCCAAACAG, reverse: ACCCCTCTTCATCTGGGCATCTG. Quantitative real-time RT-PCR of the 2 mRNAs was performed by specific gene primers using Thermal Cycler Dice Real Time (Takara Bio Inc.) according to the protocol. The primers for the 2 genes primers were as follows: C20orf85, forward: GCCATGCCAGGGAAAGGAAGAG, reverse: GAAGGGTGACGCTGATGACTTGAG; CST1, forward: AGGAGGAGGATAGGATAATCCC, reverse: TCTTTGGTGGCCTTGTTATACT. The results were normalized to β-actin and then calculated with the ΔCT method, the primers were as follows: ACTB,

forward: CCTGGCACCCAGCACAAT, reverse: GGGCCGGACTCGTCATAC. All data were expressed as the mean ± standard error of the measurement from at least three experiments.

TME Cell type enrichment abundance analysis between high-risk and low-risk groups

Components of the TME in high-risk and low-risk breast cancer samples from the TCGA cohort were assessed by applying xCell web tool (http://xcell.ucsf.edu/).The cell type enrichment score was calculated based on the TME gene expression data, the xCell tool provides 64 cell types, including immunocytes, stromal cells, stem cells, and other cells.

Construction of a ceRNA Network

MiRanda is an information resource for experimentally validated miRNA-target interactions that provides the expression profile of a miRNA and its target gene□http://www.microrna.org/microrna/home.do□. The lncRNA-miRNA interactions were predicted with the LncBase Predicted v.2 (http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2%2Findex-predicted. The Pearson correlation coefficient (PCC) for miRNA-mRNA and miRNA-lncRNA was calculated using paired miRNA, mRNA, and lncRNA expression profile data. A candidate pair of lncRNA-miRNA-mRNA was constructed based on the "ceRNA hypothesis" as follows: (i) mRNAs and lncRNAs share the same miRNAs; (ii) mRNAs and lncRNAs suggest a positive correlation when the PCC is higher than 0.3 and *P* value < 0.05; (iii) mRNAs and lncRNAs show a negative regulation with miRNA with PCC < 0 and *P* value < 0.05; and (iv) the miRNAs are aberrantly expressed in breast cancer. By integrating the lncRNA-mRNA, lncRNA-miRNA, and mRNA-miRNA pairs, the lncRNA-miRNA-mRNA ceRNA network was constructed and visualized using Cytoscape.

Statistical analysis

Student's t test was used for continuous variables, while categorical variables were compared by χ2 test. The logrank test was performed for comparing Kaplan-Meier curves between groups. The differential proportions of the TME cells in the TCGA cohort were evaluated by the Wilcoxon rank-sum test. The specificity and sensitivity of survival prediction according to the determined risk score were obtained by time-dependent receiver operating characteristic (ROC) curves, with AUC values quantified with the survival ROC package (https://cran.r-project.org/web/packages/survivalROC/index.html). We considered p <0.05 to be statistically significant.

Results

60 genes and 54 lncRNAs were identified prognosis-associated

After the first quality check by WGCNA R package, 7 tumor samples and 49 normal low-quality samples were removed from subsequent analysis, the remaining samples needed to be integrated and processed. Statistical analysis software R was used for preprocessing the differential expression analysis of microarray data. The data from the statistical results included the 753 differentially expressed genes,1315 differentially expressed lncRNAs. The volcano plot is used to show the significantly different expressions of the two sets of samples. The p-value and fold change values obtained by accurate T-test statistical analysis were used to draw the volcano map between the two groups (Figure1A, FDR < 0.05 and |log2 fold change (FC)| ≥ 2). In the volcano plot, one of the coordinates shows the negative log of p-values computed by the T-test, and the other shows the converted log of p-values compared to the two conditions. Similarly, we can also find that compared with the normal sample, the breast cancer samples down-regulated more in the differentially expressed genes and lncRNAs. Single variable Cox proportional risk regression analysis was performed to screen genes significantly associated with overall survival (OS) in the TCGA breast cancer dataset. A total of 874 genes and 3262 lncRNAs were significantly associated with prognosis. The set analysis software was used to take the intersection of the two (Venn diagram) to obtain 60 genes and 54 lncRNAs, which were not only differentially expressed between normal and breast cancer tissues, but also had prognostic differences in cancer tissues (Figure1B). Only the intersected genes and lncRNAs were selected for the subsequent bioinformatics.

In order to inquire about the potential signal pathways related to prognosis related genes in breast cancer, we screened the top30 prognosis related DEGs and analyzed them with GO and KEGG by R-packet “limma”, and the screening criteria were | lgfc | > 2, and adj. P < 0.05. GO analysis results showed that prognosis related DEGs can be enriched in basic biological processes(BP), including neurological system process, sensory perception; in cellular component, extracellular space and chromosome were enriched; in molecular function, receptor binding and protein dimerization activity were enriched. KEGG analysis showed that the top30 prognosis related DEGs were mainly enriched in systemic lupus erythematosus, cytokine-cytokine receptor interaction, salivary secretion and chemokine signaling pathway and cardiac muscle contraction pathways (Figure 1C).

Although accumulating studies have attempted to reveal the functional significance of lncRNAs, the biological roles of most lncRNAs are still unknown. Biological processes and cellular regulation networks are very complex, involving the interactions of various molecules, such as proteins, RNAs, and DNAs. We constructed an lncRNA/ mRNA co-expression network based on the correlation coefficient of lncRNAs and mRNAs and investigated the potential interaction between mRNAs and lncRNAs. The co-expression network was composed of 27 differentially expressed lncRNAs, 32 differentially expressed mRNAs and 11 network nodes (Figure 1D). The network showed that several lncRNAs (AC104211.2, RBMS3-AS3, AC002401.3, AC063919.1 and BOK-AS1) correlated with a great number of targeted mRNAs, and vice versa. This co-expression network also indicated that one lncRNA (AC104211.2) could target 15 network

nodes and one coding gene (LVRN) could target 17 network nodes. In addition, the second ranked lncRNA (RBMS3-AS3) and mRNA (KLHL33), could both target 15 network node. Taken together, these results suggested the closer inter-regulation of lncRNAs and mRNAs in breast cancer. Through the co-expression network we obtained 31 hub genes, including 15 lncRNAs (AC104211.2, RBMS3-AS3, AC002401.3, AC063919.1, BOK-AS1, AL020994.3, LINC01497, LINC02766, AC135584.1, LINC01612, AP003031.3, LINC01402, AC025280.1, AC096637.3, LINC00922) and 16 genes (LVRN, KLHL33, HIST1H2AM, PENK, SLC24A2, HIST1H2AD, HIST1H2AJ, MIR23A, MIR4269, C20orf85, CDC20B, CST2, HIST1H1T, HIST1H2AL, OR2B6, RDM1).

Preliminary Identification of Optimal Prognostic Biomarkers

LASSO logistic regression were used to screen the characteristic genes by appling the glmnet package. The 31 genes entered the LASSO regression analysis (Figure 2.A, B) and then further analyzed by Stepwise Multivariate Cox Regression analysis to establish a prognostic model for patients with breast cancer on the basis of lncRNA and gene expression levels. Finally, 3 lncRNAs and 2 genes (LINC01497, LINC02766, LINC02528, C20orf85, CST1) related to the prognosis of breast cancer were obtained. The risk scores for patients were calculated on the basis of the relevant RNA expression level and risk coefficient of each gene, and patients were categorized as low or high risk according to the optimal cut-off (Figure 2C). The AUC of risk score was 0.634, which proved that the Cox model had an ability to predict prognosis and was an independent prognostic indicator (Figure 2D). Scatter plot was drawn to show gene expression profiles in high-risk and low-risk groups (Figure2f). Kaplan-Meier curves of OS in all patients with breast cancers based on 3 lncRNAs and 2 genes expression showed that the survival time of patients with low-risk score was significantly longer than that of patients with high-risk score (Figure 2E).

The 5 biomarkers were independent prognostic indicators

In an attempt to confirm the independent prognostic impact of individual lncRNA or gene, we performed univariate COX regression of the 5 screened variables, the results showed that the 5 biomarkers coul be independent prognostic indicator (Figure 3A). Calculated by multivariate Cox regression model, hazard ratio with 95% confidence interval (95% CI) for independent 3 lncRNAs and 2 genes signature were shown in forest plot. The forest map results showed that the genes with HR>1 (ENSG00000237560(LINC01497), ENSG00000229484(LINC02766)) are considered to be dangerous genes, while the genes with HR<1(ENSG00000124237(C20orf85), ENSG00000170373 (CST1), ENSG00000226004(LINC02528)) to be a protective gene (Figure 3B).

Composition differences of TME between high and low risk breast cancer groups and their associations with prognosis

Of all breast cancer samples, high and low risk groups were eligible based on xCell analysis. The results showed that there were 44 cell types differently expressed between high and low risk groups, and the immune score and stroma score were also different between the two groups (Figure 4A, $p<0.05$). The two most differently expressed components of the TME in breast cancer tissues were Epithelial cells and HSCs. Specifically, the proportions of Adipocytes, Endothelial cells, HSCs, Fibroblasts were significantly lower in low risk score tissues compared with the high risk score tissues, while the proportions of M1 macrophages, MSCs, Th2 cells were significantly higher (Figure 4B). Correlations among the components ranged from weak to moderate. Obviously, Astrocytes showed highly positive correlations with CD8+ Tcm and MSCs, Myocytes showed highly positive correlations with Mast cells, Macrophages M2 and Osteoblast (Figure 4C). Then single variable Cox proportional risk regression analysis was performed to screen cells types significantly associated with overall survival (OS). The results showed that there were 13 differently expressed cell types between high and low risk groups were associated with survival, of which, aDC, CLP, Melanocytes, NKT, pDC, Tgd.cells were protection factors, while Endothelial cells, Hepatocytes, ly Endothelial cells, mv. Endothelial cells, Neurons, Neutrophils, Preadipocytes were associated with poor prognosis (Figure 5), indicating that the TME compositions have high sensitivity and specificity to predict the prognosis of breast cancer patients.

Validation of the 5 indicators expressions Using RT-PCR

To verify the reliability of the aberrant lncRNAs and mRNAs found in the TCGA database, a RT-PCR assay was used to detect the expression levels of the 3 lncRNAs and 2 mRNAs in 25 paired tumor tissues and paracancer tissues from patients with breast cancer. The results indicated that the levels of LINC01497 and LINC02766 were decreased in the tumor tissues, and the LINC02528,c20orf85 and CST1were upregulated(Figure 6A,n=10,tumor vs normal, Student's t test,*$p<0.05$). The results were consistent with the expressions in TGCA BRCA data.

Construction of the breast cancer-specific ceRNA Network

The biological roles of most lncRNAs are still unknown, accumulating studies have attempted to reveal the functional significance of lncRNAs on gene by modulating miRNAs. To establish the ceRNA network, we obtained the miRNAs that might interact with both the 3lncRNAs and 2 genes. Based on the results of the online prediction, by integrating them, a ceRNA network containing 3 lncRNAs, 2 mRNAs, and 158 miRNAs was finally constructed (Figure 6B). A total of 286 pairs of miRNA-lncRNA interactions containing 147 miRNAs and 3 lncRNAs, 70 pairs of miRNA-mRNA interactions containing 52 miRNAs and 2 mRNAs were contained in the ceRNA network. The degree of nodes in the ceRNA network was calculated and red color presents a highest degree, and purple presents a lowest degree.

Discussion

Worldwide, the incidence of breast cancer is increasing and high risk of recurrence and the high mortality raise concerns to women health. Great efforts are put by clinicians and researchers and progressions are seen in early detection, diagnosis, and treatments of breast cancer over the years，which suggests a benefit from the combination of early detection and more effective comprehensive treatment of local disease with surgery, radiation therapy, endocrine therapy, and systemic treatment with chemotherapy and so on[2]. Nevertheless, early recurrence, distant metastasis and drug resistance are still commonly seen, which hold threads to the prognosis of breast cancer patients and mount challenges for clinicians[12-14]. Even though patients with the same molecular subtype of breast cancer receive the same therapy, they have different outcomes[15, 16]. Hence, further researches were urgently needed to unravel the molecular mechanism underlying and discovering valuable prognostic biomarkers for breast cancer survival and thus to provide guidance for individualized treatment.

LncRNAs have been shown to be involved in mammary gland development, as well as breast cancer evolution[17, 18]. More and more studies have confirmed that abnormal lncRNA expression is related to the progress of breast cancer and can be used as markers for cancer detection and prognosis, as well as targets for breast cancer treatment[19-21]. Compared with protein-coding genes, lncRNAs have significant advantages as diagnostic and prognostic biomarkers, the association between lncRNAs signature and breast cancer patients' survival has been assessed in various studies[22]. Many lncRNAs, such as lncRNA MALAT1,whose levels were found inversely correlate with breast cancer progression and metastatic ability, was related with the breast cancer patients' survival [23, 24].Others such as HOTAIR, XIST,H19 U79277,AK024118, BC040204 and AK000974 have been shown to be aberrantly expressed in breast cancer and cell lines and have been recognized predictive of breast cancer patient survival[20, 21]. Besides, a recent study has also identified 3 lncRNA genes (LINC00324, PTPRGAS1 and SNHG17) associated with tumor histologic grade, as well as clinical outcomes[25]. However, accurate prognostic markers are still lacking. In this study, we attempted to use lncRNA data, mRNA data and clinical follow-up data from the tumor genome atlas to constructed a prognosis-specific lncRNAs/mRNAs co-expression network based on correlation analysis to identify breast cancer prognostic factors and provide potential therapeutic targets for treatment.

In this study, breast cancer-specific genes and lncRNAs were identified through bioinformatics analysis of tens of thousands of candidate RNAs, and by applying single variable Cox proportional risk regression analysis to select the genes that were related with the survival. Then the ones that were both abnormally expressed and related with survival were obtained. Through lncRNA/mRNA co-expressed network construction, 15 lncRNAs and 16 mRNA hub genes were gained. The genes entered to LASSO and multivariate Cox proportional hazard regression analysis, finally, 3 lncRNAs (LINC01497, LINC02766, LINC02528) and 2 mRNAs (C20orf85, CST1) were selected as prognosis predictive genes. According to the median risk score of the 5 genes, patients were divided into high-risk group and low-risk group, patients with low-risk score was significantly longer than that of patients with high-risk score. The AUC of risk score was significantly large, which proved that the Cox model had better ability to predict prognosis. Besides, through univariate COX regression analysis of the 5 screened variables, we confirmed that the 5 screened variables can be independent prognostic indicator. The three lncRNAs have not been reported in

previous studies so far and it was the first time that they were identified as prognostic indicators. Cystatin SN (CST1),one of the 5 screened prognostic indicators, is a specific inhibitor of cysteine proteases with the activity of protein degradation and was reported associated with a diversity of diseases and facilitates the development and progression of cancer cells[26]. Various research have inferred that CST1 acts as an imperative role in tumorigenesis and tumor suppressors[27, 28]. For instance, upregulation of CST1 promotes gastric carcinoma cells proliferation and inhibits cathepsin activity[29]. High CST1 expression was also reported by previous studies to be linked to poor survival in colorectal cancer[30], pancreatic cancer[31], and non-small cell lung cancer patients[32]. In this study, our results showed that CST1 was upregulated in breast cancer tissues, which was consistent with previous reports. However, its high expression was related with a longer survival time. In order to explore the reason, we analyzed the CST1 expression in TCGA BRCA subclasses data and found CST1 was upregulated most in luminal subtype and second in Her-2 positive subtype, while has a least increase in TNBC subtype. This implied that CST1 could serve as a feasible prognostic factor that related with tumor types and may be a new biomarker in breast cancer. C20orf85, also known as LLC1, was firstly found present in the normal lung epithelium but absent or downregulated in most primary nonsmall cell lung cancers and lung cancer cell lines, but was not studied before in breast cancers. Interestingly, this is the first time to identify c20orf85 could act as a prognostic marker for breast cancer. Besides, to confirm the accuracy of the analysis results, we did RT-PCR experiments and the results were consistent with the bioinformatics analysis results, indicating our method to be reliable. However, what's the correlation between the 3 lncRNAs and 2 mRNAs as they were finally identified prognostic indicators. The studies on the participation of lncRNAs in gene regulation through the ceRNA mechanism, as in through the lncRNA/miRNA/mRNA axis, in the pathogenesis, progression, and metastasis accumulated over the past few years[33]. The competing endogenous RNA (ceRNA) hypothesis was first proposed by Salmena and colleagues, lncRNAs can be enhancers, scaffolds, "sponges" that bind several miRNAs, or even precursors of some miRNAs to mutually regulate their expression or functions[34]. So a dysregulated lncRNA-associated ceRNA network was constructed based on the 3 lncRNA, and the 2 mRNA and the miRNA expression profiles by using an integrative computational method, and finally construct a ceRNA network, which initially reveals a possible regulatory mechanism between lncRNAs and mRNAs.

Increasing evidence demonstrated the importance of the tumor microenvironment (TME)，a mixture that consists of mesenchymal cells, tumor-infiltrating immune cells (TIICs), endothelial cells, extracellular matrix molecules and inflammatory mediators in the tumor development. Structural components of the TME are mainly resident stromal cells and recruited immune cells, while there was compelling evidence for the role of stromal cell contributing to tumor angiogenesis and extracellular matrix remodeling, but perhaps it is still not fully understood[35]. Collaborative interactions between cancer cells and their supporting cells contributed to the malignant phenotypes of cancer, such as immortal proliferation, resisting apoptosis, and evading immune surveillance, therefore, the TME significantly influences therapeutic response and clinical outcome in cancer patients[36]. In the TME, TIICs constitute the major type of non-tumor components reported to be valuable for prognostic assessment in OS[37]. Prognostic or predictive biomarkers, associated with tumor immune microenvironment, may have great prospects in

guiding patient management, identifying new immune-related molecular markers, establishing personalized risk assessment of breast cancer[38]. Thus, improving immunotherapy efficacy in breast cancer by systematically assessing the TME's immune properties and determining components of TME distribution in high and low risk group patients is of prime importance. Cell type Identification by xCell was used to assess the levels of components of TME in large amounts of heterogeneous samples [14]. Cell has been successfully applied for identifying components of TME cells landscapes and their associations with prognosis in neuroblastoma and triple-negative breast cancer[39, 40]. Large-scale public data with gene expression and clinical information, complete biological databases, and sophisticated high-throughput data analysis methods together provide opportunities for identifying broader prognostic features in breast cancer biology. In this study, to understand the dynamic modulation of the immune and stromal components in TME in high and low risk group patients, we applied xCell computational methods to compare the TME components in high-risk and low-risk groups. The results showed that the proportions of Adipocytes, Endothelial cells, HSCs, Fibroblasts were significantly lower in low risk score tissues compared with the high risk score tissues, while the proportions of M1 macrophages, MSCs, Th2 cells were significantly higher. Subsequently, optimal TME cells were selected as prognostic indictors by univariate Cox regression analyses. It has been demonstrated that 13 types: Neutrophils, NKT, pDC, ly Endothelial cells, Melanocytes, Endothelial cells, Tgd cells, Hepatocytes, mv Endothelial cells, Neurons, Preadipocytes, CLP, aDC were differently expressed both between tumor and normal groups and high risk and low risk groups, suggesting the proportions of TME were associated with the outcome in breast cancer patients. It adds strong evidence that tumor-associated immune genes may become potential targets for cancer therapy.

Conclusions

Using bioinformatics analysis both in lncRNA and mRNA levels, we focused on survival-related genes of breast cancer based on large data of real samples, and screened five potential prognostic targets. Although the prognostic markers identified in our current study may still need further researches and validations, they may provide different insights into prognostic monitoring for breast cancer.

Abbreviations

LASSO: least absolute shrinkage and selection operator

GO: Gene Ontology

Kyoto Encyclopedia of Genes and Genomes

PCC: Pearson correlation coefficient

TME: Tumor microenvironment

TIICs: tumor-infiltrating immune cells

Declarations

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Funding

This work was supported by grants from the National Natural Science Foundation of China (81873730) and the Jiangsu Women and Children Health Key Discipline Program (FXK201758).

Author contributions

JW researched and wrote the manuscript, QD did the experiments and analyzed data. FB prepared the figures. G-QJ designed and supervised the experiments, and edited the manuscript. All authors discussed the results and commented on the manuscript. JW and QD contributed equally to the work.

Acknowledgements

Not applicable.

Ethics approval and consent to participate

Study approval was obtained from independent ethics committees at the Second Affiliated Hospital of Soochow University. All subjects involved in this study signed informed consent forms. The privacy of patients involved were protected.

References

1. Howell A, Anderson AS, Clarke RB, Duffy SW, Evans DG, Garcia-Closas M, Gescher AJ, Key TJ, Saxton JM, Harvie MN. Risk determination and prevention of breast cancer. *Breast cancer research : BCR* 2014, **16**(5):446-446.
2. Colditz GA, Bohlke K. Priorities for the primary prevention of breast cancer. *Ca A Cancer Journal for Clinicians* 2014, **64**(3).
3. Fragomeni SM, Sciallis A, Jeruss JS. Molecular Subtypes and Local-Regional Control of Breast Cancer. *Surg Oncol Clin N Am* 2018, **27**(1):95-120.
4. Greenlee H, DuPont-Reyes MJ, Balneaves LG, Carlson LE, Cohen MR, Deng G, Johnson JA, Mumber M, Seely D, Zick SM *et al*. Clinical practice guidelines on the evidence-based use of integrative therapies during and after breast cancer treatment. *CA: a cancer journal for clinicians* 2017, **67**(3):194-232.
5. Zhu Y, Luo M, Brooks M, Clouthier SG, Wicha MS. Biological and clinical significance of cancer stem cell plasticity. *Clinical and Translational Medicine* 2014.
6. Gibb, Ewan, A., Brown, Carolyn, J., Lam, Wan, L. The functional role of long non-coding RNA in human carcinomas. *Molecular Cancer* 2011.
7. Muhammad, Al-Hajj, Max, S., Wicha, Adalberto, Benito-Hernandez, Sean, J., Morrison: From the Cover. Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America* 2003.
8. Askarian-Amiri ME, Seyfoddin V, Smart CE, Wang J, Ji EK, Hansji H, Baguley BC, Finlay GJ, Leung EY. Emerging Role of Long Non-Coding RNA SOX2OT in SOX2 Regulation in Breast Cancer. *Plos One* 2014, **9**.
9. Nguyen NP, Almeida FS, Chi A, Nguyen LM, Cohen D, Karlsson U, Vinh-Hung V: Molecular biology of breast cancer stem cells. Potential clinical applications. *Cancer Treatment Reviews* 2010, **36**(6):485-491.
10. Richards EJ, Zhang G, Li ZP, Permuthwey J, Challa S, Li Y, Kong W, Dan S, Bui M, Coppola D: Long non-coding RNAs regulated by TGFβ. lncRNA-HIT mediated TGFβ-induced epithelial to mesenchymal transition in mammary epithelia. *Journal of Biological Chemistry* 2015, **290**(11):6857-6867.
11. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010, **33**(1):1-22.
12. Freddie, Bray, Jacques, Ferlay, Isabelle, Soerjomataram, Rebecca, Siegel, Lindsey. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca A Cancer Journal for Clinicians* 2018.
13. Robert, Burton, Robin, Bell. The global challenge of reducing breast cancer mortality. *Oncologist* 2013.
14. McArthur, Heather. Breast Cancer Brain Metastasis: An Ongoing Clinical Challenge and Opportunity for Innovation. *Oncology* 2016.
15. Yang Y, Im SA, Keam B, Lee KH, Kim TY, Suh KJ, Ryu HS, Moon HG, Han SW, Oh DY *et al*. Prognostic impact of AJCC response criteria for neoadjuvant chemotherapy in stage II/III breast cancer patients:

breast cancer subtype analyses. *BMC cancer* 2016, **16**:515.

16. Bergen ES, Tichy C, Berghoff AS, Rudas M, Bartsch R. Abstract P2-08-17: Prognostic impact of breast cancer subtypes in elderly patients. *Cancer Research* 2016, **76**(4 Supplement):P2-08-17-P02-08-17.
17. Hansji H, Leung EY, Baguley BC, Finlay GJ, Askarian-Amiri ME. Keeping abreast with long non-coding RNAs in mammary gland development and breast cancer. *Frontiers in genetics* 2014, **5**:379.
18. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. *British journal of cancer* 2013, **108**(12):2419-2425.
19. Arun G, Spector DL. MALAT1 long non-coding RNA and breast cancer. *RNA biology* 2019, **16**(6):860-863.
20. Meng J, Li P, Zhang Q, Yang Z, Fu S. A four-long non-coding RNA signature in predicting breast cancer survival. *Journal of experimental & clinical cancer research : CR* 2014, **33**(1):84.
21. Soudyab M, Iranpour M, Ghafouri-Fard S. The Role of Long Non-Coding RNAs in Breast Cancer. *Archives of Iranian medicine* 2016, **19**(7):508-517.
22. Kosir MA, Jia H, Ju D, Lipovich L. Challenging paradigms: long non-coding RNAs in breast ductal carcinoma in situ (DCIS). *Frontiers in genetics* 2013, **4**:50.
23. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010, **464**(7291):1071-1076.
24. Kim J, Piao HL, Kim BJ, Yao F, Han Z, Wang Y, Xiao Z, Siverly AN, Lawhon SE, Ton BN *et al*. Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nature genetics* 2018, **50**(12):1705-1715.
25. Zhao W, Luo J, Jiao S. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Scientific reports* 2014, **4**:6591.
26. Löser R, Pietzsch J. Cysteine cathepsins: their role in tumor progression and recent trends in the development of imaging probes. *Frontiers in chemistry* 2015, **3**(70):37.
27. Pi?Lar A, Nanut MPI, Kos J. Lysosomal cysteine peptidases – Molecules signaling tumor cell death and survival. *Seminars in Cancer Biology* 2015, **35**:168-179.
28. Vasiljeva O, Turk B. Dual contrasting roles of cysteine cathepsins in cancer progression: Apoptosis versus tumour invasion. *Biochimie* 2008, **90**(2):380-386.
29. Choi EH, Kim JT, Kim JH, Kim SY, Song EY, Kim JW, Kim SY, Yeom YI, Kim IH, Lee HG. Upregulation of the cysteine protease inhibitor, cystatin SN, contributes to cell proliferation and cathepsin inhibition in gastric cancer. *Clinica Chimica Acta* 2009, **406**(1-2):45-51.
30. Nakajima. Identification of Cystatin SN as a novel tumor marker for colorectal cancer. *International Journal of Oncology* 2009, **35**(01):33-40.
31. Jiang J, Liu HL, Liu ZH, Tan SW, Wu B. Identification of cystatin SN as a novel biomarker for pancreatic cancer. *Tumor Biology* 2015, **36**(5):3903-3910.

32. Cao X, Li Y, Luo RZ, Zhang L, Zhang SL, Zeng J, Han YJ, Wen ZS. Expression of Cystatin SN significantly correlates with recurrence, metastasis, and survival duration in surgically resected non-small cell lung cancer patients. *Rep* 2015, **5**:8230.
33. Anna SC, Yumi K, Yusuke Y, Fumitaka T, Takahiro O. The Emerging Roles of Long Non-coding RNA in Cancer. *Cancer Science* 2018, **109**.
34. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011, **146**(3):353-358.
35. Bussard KM, Mutkus L, Stumpf K, Gomez-Manzano C, Marini FC. Tumor-associated stromal cells as key contributors to the tumor microenvironment. *Breast cancer research : BCR* 2016, **18**(1):84.
36. Wood SL, Pernemalm M, Crosbie PA, Whetton AD. The role of the tumor-microenvironment in lung cancer-metastasis and its relationship to potential therapeutic targets. *Cancer Treatment Reviews* 2014, **40**(4):558-566.
37. Wang C, Zhou X, Li W, Li M, Tu T, Ba X, Wu Y, Huang Z, Fan G, Zhou G. Macrophage migration inhibitory factor promotes osteosarcoma growth and lung metastasis through activating the RAS/MAPK pathway. *Cancer Letters* 2017:S0304383517303920.
38. Li X, Chen Y, Liu X, Zhang J, He X, Teng G, Yu D. Tim3/Gal9 interactions between T cells and monocytes result in an immunosuppressive feedback loop that inhibits Th1 responses in osteosarcoma patients. *International Immunopharmacology* 2017, **44**(Complete):153-159.
39. Zhong X, Zhang Y, Wang L, Zhang H, Liu H, Liu Y. Cellular components in tumor microenvironment of neuroblastoma and the prognostic value. *PeerJ* 2019, **7**:e8017-e8017.
40. Oshi M, Asaoka M, Tokumaru Y, Angarita FA, Yan L, Matsuyama R, Zsiros E, Ishikawa T, Endo I, Takabe K. Abundance of Regulatory T Cell (Treg) as a Predictive Biomarker for Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer. *Cancers (Basel)* 2020, **12**(10):3038.

Figures

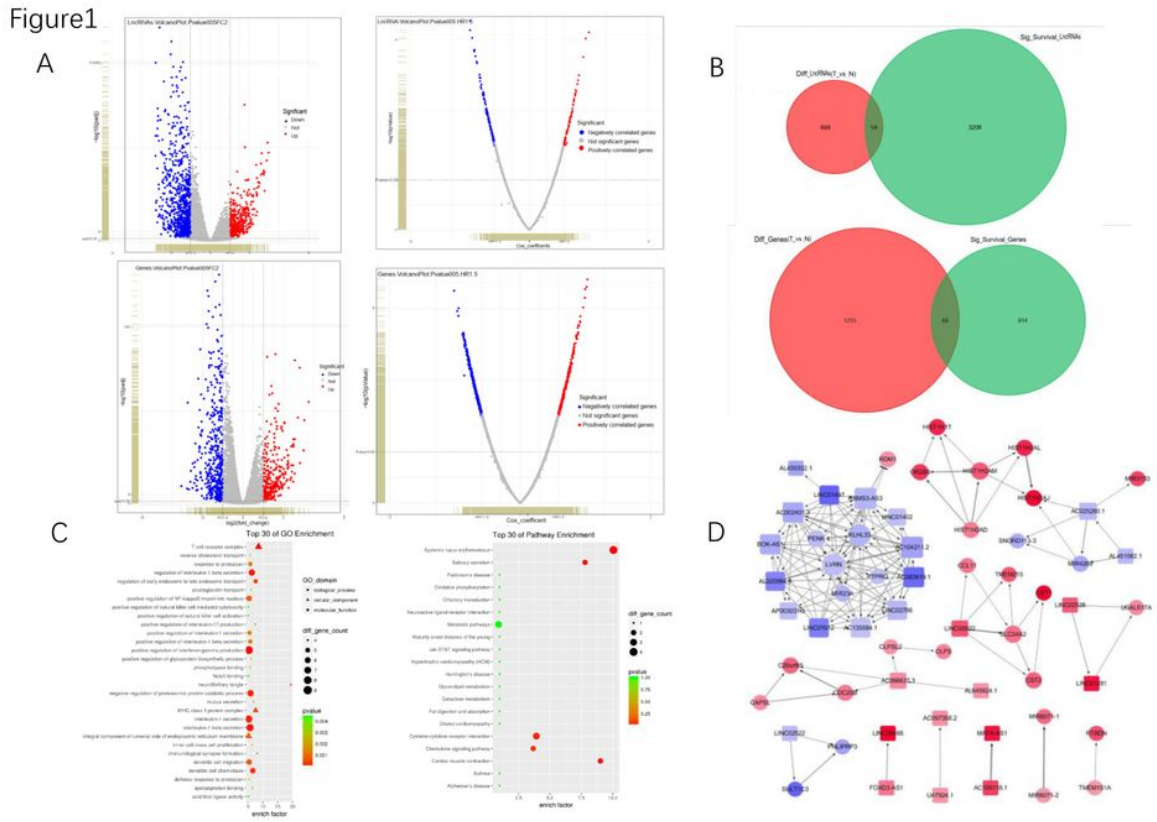

Figure 1

Volcano plots（A） , Venn diagrams(B) , GO and KEGG enrichment (C) and lncRNA-mRNA interaction network (D). A. left: Blue represents the fold change value of DEGs > 2, blue represents the fold change value of DEGs <-2. p < 0.05 was considered statistically significant. Right: Red represents DEGs with positively correlated; blue represents DEGs with negatively correlated, gray represents DEGs IHRI<1.5, p < 0.05 was considered statistically significant; B. Venn plots showing common survival-related and differently expressed lncRNAs (up) and genes (down); C. GO and KEGG enrichment analysis for top30 differently expressed survival related genes, terms with p < 0.05 were believed to be enriched significantly. D. Interaction network constructed with the nodes with interaction confidence value >0.95.

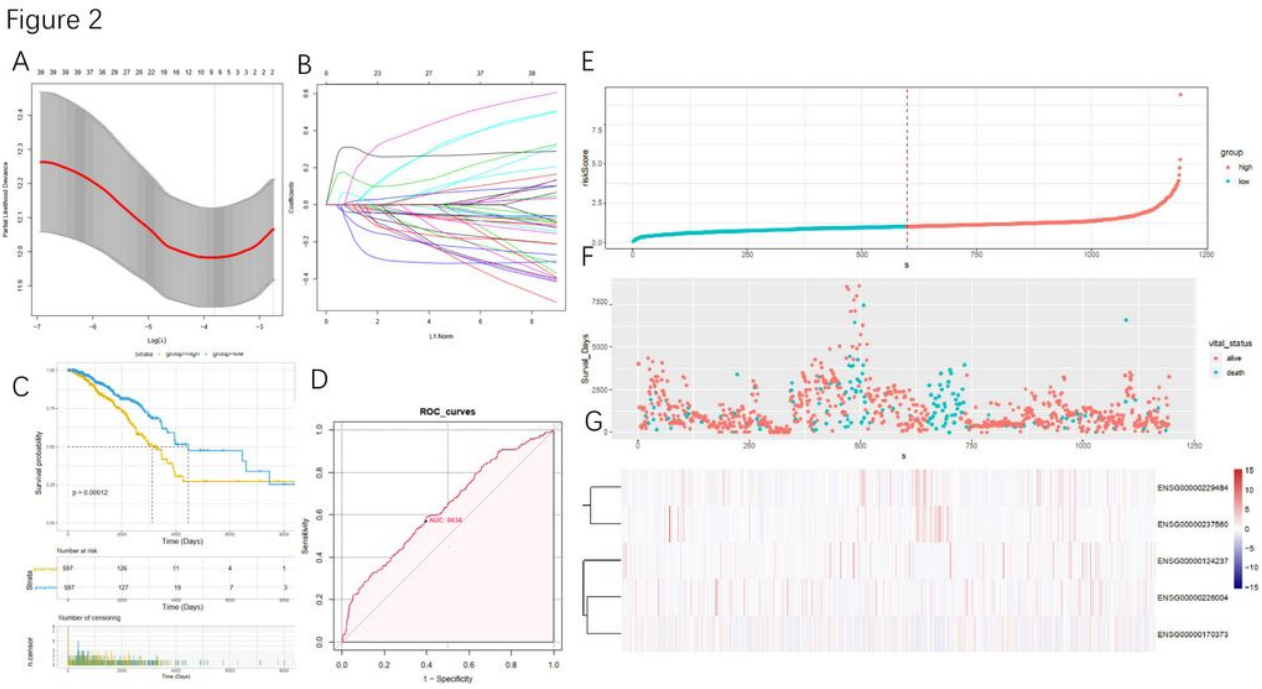

Figure 2

LASSO feature selection and distribution of risk score, survival time and gene expression panel included in the risk model. A. Selecting the best parameters for STAD in the LASSO model. B. LASSO coefficient profiles of the most useful prognostic genes. C. Kaplan-Meier curves in all patients with breast cancer based on risk score the patients in low-risk group has a longer survival time; D. Survival prediction based on the risk score, determined by time-dependent ROC curve; E. the distribution of risk score; F. the plot of death status; G. heatmap of 3 lncRNAs and 2 mRNAs expression.

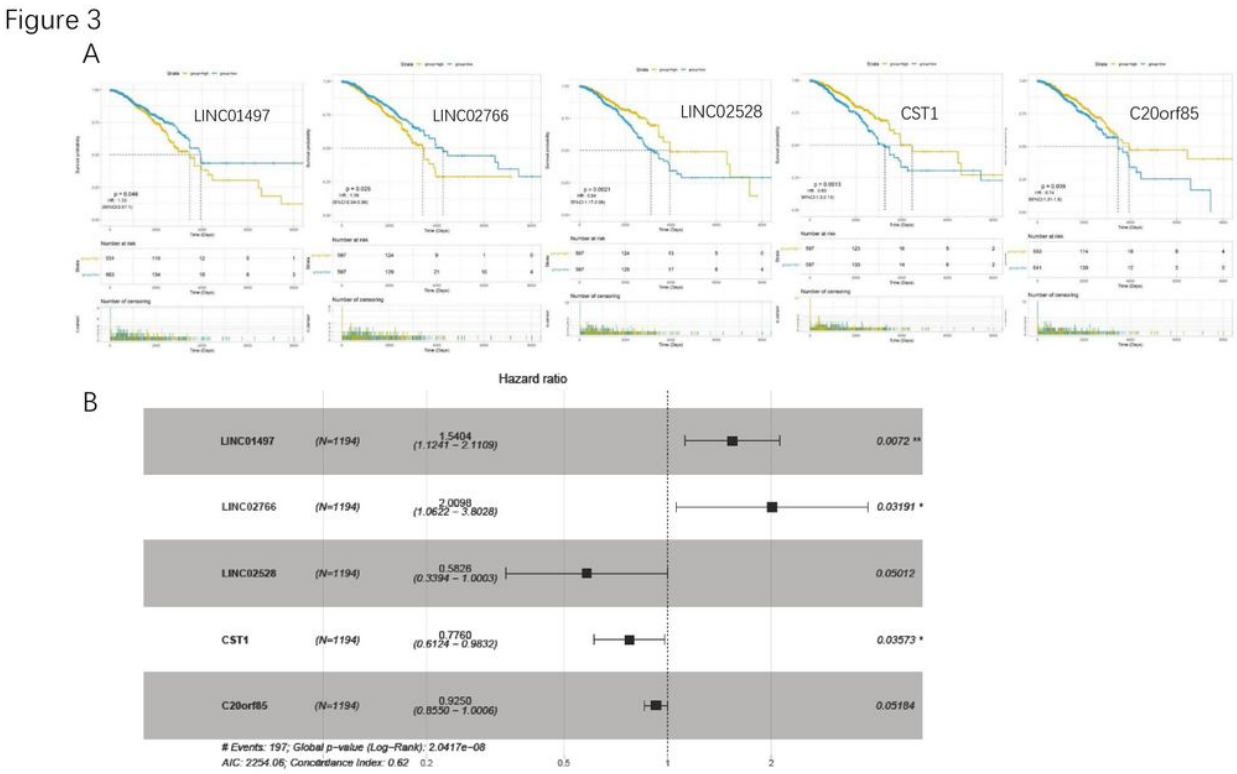

Figure 3

Univariate COX regression analysis with the 3lncRNAs and 2 mRNAs (A) and the forest map of the five potential biomarkers(B). A. Kaplan-Meier curves for the 5 biomarkers. B. Forest map: Calculated by multivariate Cox regression model, hazard ratio with 95% confidence interval (95% CI), p <0.05 was considered statistically significant.

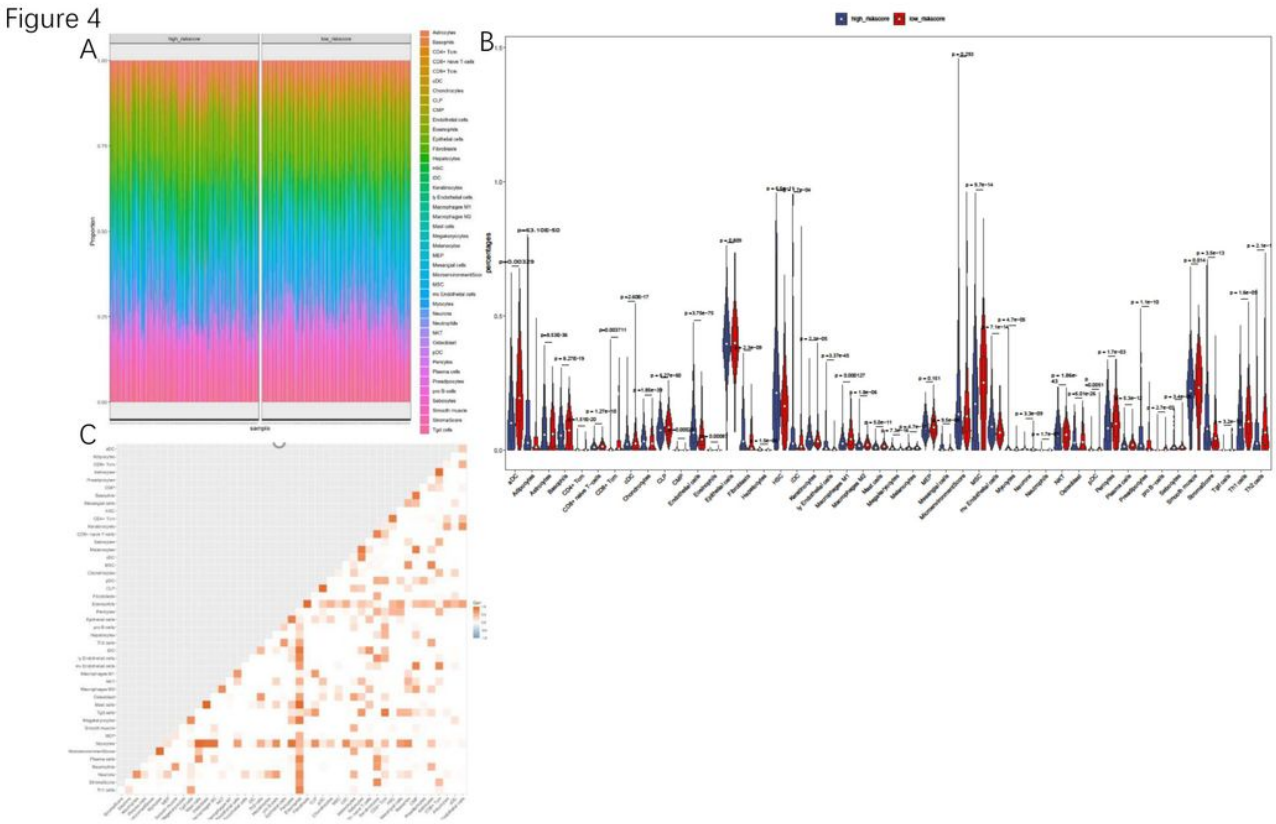

Figure 4

Heatmap（A）, Boxplot and violinplot of Different immune celltypes between low and high risk score breast cancer samples(B) and correlation matrix(C) of TME cells proportions. A. Heatmap of TME cells proportions, the expression level of TME cells with high to low levels are shown in orange, green, blue, pink respectively; B. Horizontal and vertical axes respectively represent immune celltypes and relative percentages. Blue and red colors represent high and low risk score breast cancer samples, respectively. Data were assessed by the Wilcoxon rank-sum test; C. Correlation matrix of TME cells proportions. Horizontal and vertical axes both represent TME cells. TME cells with higher, lower correlation levels are shown in red, blue respectively.

Figure 5

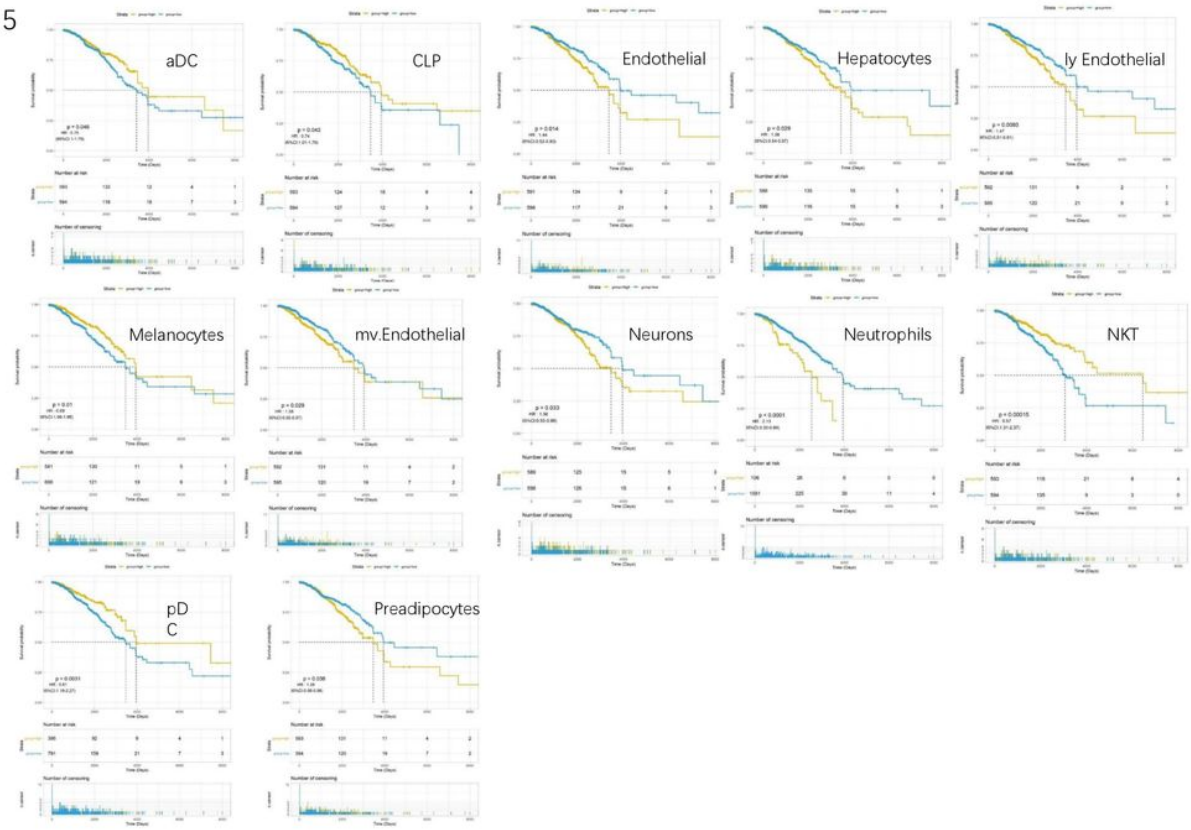

Figure 5

Overall survival of the 13 TME subtypes in breast cancer based on Kaplan-Meier plotter. The patients were stratified into high-level group and low-level group according to median expression.

Figure 6

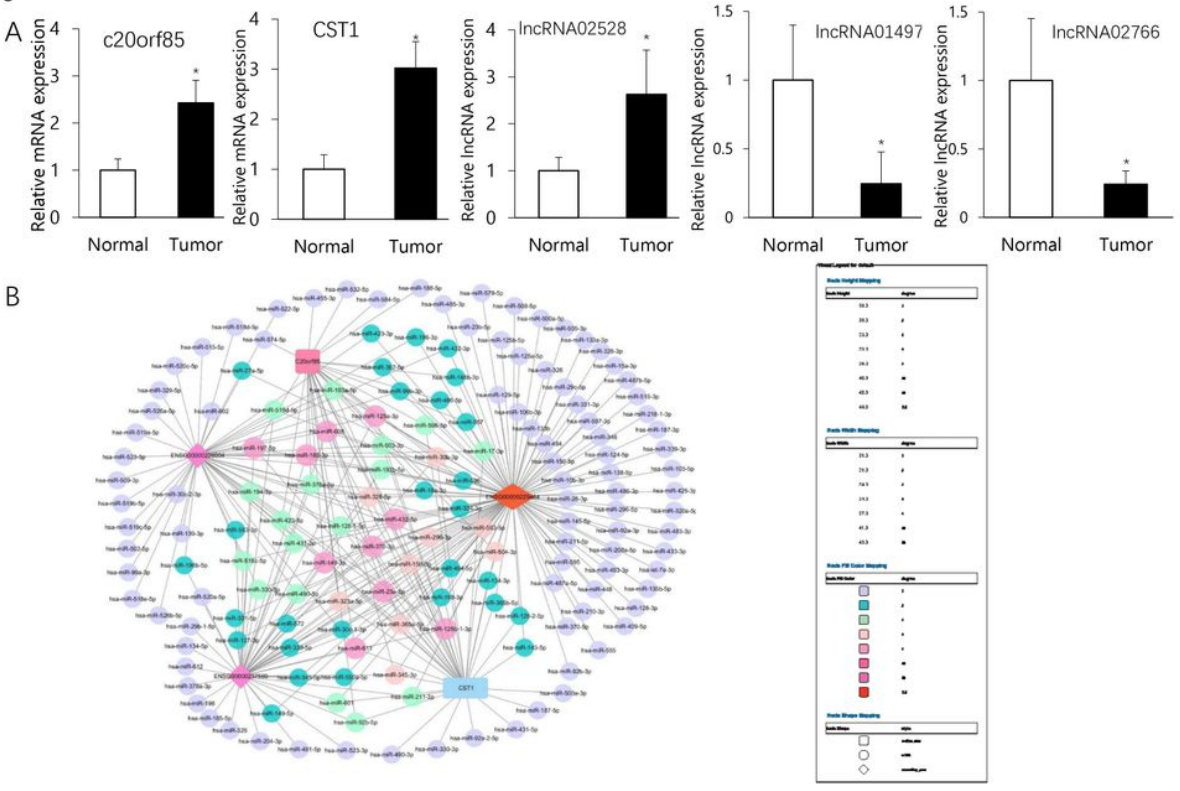

Figure 6

Validation by qPCR (A) and a ceRNA network map(b) for of the 5 biomarkers. A. The relative expression levels of the 3 lncRNAs and 2 mRNAs were detected by qPCR in the tumor tissues and paired adjacent normal breast tissues from 25 patients with breast cancer. The data are presented as the relative expression level in tumor tissues compared with normal control tissues. $*p< 0.05$ vs. normal. The data are expressed as the mean ± standard error of measurement from at least three experiments. b. Graphical view of the lncRNA-miRNA-mRNA network in breast cancer. The color of the nodes represents the power of the interrelation among the nodes. In the network, genes are shaped in rectangle, and lncRNAs are shaped in diamond and miRNAs are shaped in round. The more edges a molecular has, the more genes/miRNAs/lncRNAs that connect to it and the more central a role it plays within the network.