Construction of a Novel 3-lncRNA Risk Score System for the Prognostic Prediction of Triple-Negative Breast Cancer

Chao Yuan

Kunming University of Science and Technology

Hongjun Yuan

Kunming University of Science and Technology

Li Chen

Kunming University of Science and Technology

Miaomiao Sheng

Kunming University of Science and Technology

Wenru Tang (✉ tangkmust_tougao@163.com)

Kunming University of Science and Technology <https://orcid.org/0000-0002-0769-6448>

Primary research

Keywords: Triple-negative breast cancer, Long non-coding RNA, Biomarker, WGCNA, GEO, TCGA, Survival analysis

Posted Date: August 5th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-52497/v1>

License: (cc) (i) This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Triple-negative breast cancer (TNBC) is an essential type of breast cancer (BC). Compared with other molecular subtypes of BC, TNBC has the features of fast tumor increase, quick recurrence and natural metastasis. It is more urgent to establish a comprehensive evaluation system containing multiple biomarkers than single parameter.

Methods We conduct a bioinformatics analysis on 13 BC expression datasets from the Gene Expression Omnibus (GEO), which covered 2950 samples. We took 3484 genes with a more significant difference between TNBC and normal-like candidate genes for weighted correlation network analysis (WGCNA). A total of 54 genes were chosen as hub genes with great connectivity with the TNBC significant module. Based on The Cancer Genome Atlas (TCGA) data, we identify the best prognostic three lncRNA. Multivariate Cox regression was used to construct a 3-lncRNA risk score model. We evaluated prognostic capacity using time-dependent subject operating characteristics (ROC) and Kaplan-Meier (KM) survival analysis. The predictive power of the model was demonstrated by the time-dependent ROC spline and Kaplan-Meier spline. At the same time, it also shows good predictive ability in the validation set. Ultimately, Functional enrichment analysis of hub genes and three lncRNAs were offered to advise the possible biological pathways.

Results The construct LNC00337, DEPCE-AS1, DDX11-AS1 multi-factor risk scoring model was meaningfully associated with the prognosis of TNBC patients. Through survival analysis, the risk score efficiently divided the patients into high-risk groups with poor overall survival. The time-dependent ROC curve revealed that the model presented robust in predicting survival over the first 3 years. The validity of the model in the validation set is also verified. Finally, functional enrichment analysis proposed some biological pathways that may be correlated to the tumor.

Conclusions In our study, we established a lncRNA-based model to prognosticate the prediction of TNBC, which might afford a strong prognosis estimate tool to help therapy policy-making in the clinic.

Introduction:

Breast cancer (BC) is a frequently diagnosed malignancy and the leading cause of cancer death among females worldwide [1]. Triple-negative breast cancer (TNBC) refers to breast cancer that does not express the genes for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her2/neu), constitutes 15–20% of all breast cancers [2, 3]. TNBC in young women with large tumors, high lymphatic metastasis rate, and high clinical stage. Thus, there is a momentous crucial in establishing a strong prognosis estimate tool for predicting TNBC and making conservative judgments based on clinical features.

Biomarkers used to predict the occurrence of cancer range from clinical features, endogenous substances and histopathological characteristics of tumors to specific mutated genes[4]. For example, the tumor node metastasis (TNM) classification system is most widely used to estimate prognosis and guide

treatment in patients with cancer[5]. In addition, studies have shown that some single biomarkers have been closely associated with the occurrence and development of TNBC, such as STAT3[6], miR-105/93-3p[7], and CYPOR[8]. However, due to the features fast tumor increase, quick recurrence and natural metastasis [9], it is challenging to predict TNBC patients' survival with a single parameter. Therefore, establishing a comprehensive prognostic assessment mode, including multiple biomarkers, is the best way to increase prediction power.

In this work, we constructed a model to predict the OS of TNBC patients based on a variety of prognostic-related lncRNAs. We reused publicly available breast cancer datasets from the Gene Expression Omnibus (GEO), and differentially expressed genes (DEGs) were screened. These DEGs were then adopted to screen gene modules highly coordinated and closely related to TNBC through weighted correlation network analysis (WGCNA). Three seldom reported hub genes, LINC00337, DEPDC1-AS1, and DDX11-AS1, were selected to construct the prognosis scoring model. The risk score was calculated by multiplying the multivariate Cox factor by the gene expression. Training sets and validation sets were conducted to validate the risk scoring model. The risk scoring model was estimated by time-dependent ROC analysis. Finally, the potential biological pathways of hub genes and 3 lncRNAs were identified by functional enrichment analysis.

Materials & Methods

Data resource

The 13 BC expression datasets were downloaded from GEO, including GSE20685, GSE21653, GSE12276, GSE42568, GSE58812, GSE102484, GSE27830, GSE45827, GSE76124, GSE65194, GSE43365, GSE36771 and GSE31448. The datasets were reviewed manually to fulfill the following criteria, including, (☒) Studies involved in selection samples >100 only, (☒) Studies without any drug treatment. The datasets were based on the GPL570 platform.

All the available BC mRNA datasets were searched for the TCGA data portal and obtained using the R package TCGAbiolinks [10]. Clinical information was downloaded and extracted to molecular subtypes of breast cancer and statistical analyses. No approval from the ethics committee was needed because all the information was required from the TCGA database (<https://portal.gdc.cancer.gov/>).

Construct feature DEGs

The 13 series have good consistency after background correction, standardized, and summarization by RMA arithmetic. We also annotated the gene biotypes of these microarray probes by the R package biomaRt [11]. All breast samples were classified into different molecular characteristics. The PAM50 classifier was used to divided BC samples of Basal-like, Luminal A, Luminal B, Her2, and Normal-like molecular characteristics by the R package geneFu [12]. Among them, Basal-like is also called TNBC. The limma package was used to analyze the DEGs in the other four subtypes, respectively, compared to the

normal-like subtypes of BC [13]. DEGs were obtained according to the criteria: adjusted top10,000. The feature DEGs with specific differences between Basal-like and Normal-like were obtained by removing the same gene probes as the other three DEGs sets.

WGCNA co-expression network construction

WGCNA analyzed these feature DEGs, and pair-wise correlation matrices were constructed[14]. Using average cascading hierarchical clustering, we classified the genes with high absolute correlation according TOM-based difference measure and the gene modules with the minimum size of 25. A heat map is drawn to show the correlation and independence of each module. Then, using correlation analysis, the most closely related modules with TNBC were selected for further analysis.

Identification and validation of Hub Genes

Hub genes usually refer to characteristic genes that are highly associated with other genes in the module. The module connectivity of each gene in relevant modules was characterized by the absolute value of module membership (MM) after the module was associated with the BC subtype. We calculated the fabs of gene significance (GS), which described the Pearson's correlation between a given gene and the molecular characteristics of BC. The selection criteria for hub genes were based on absolute GS> 0.2, absolute MM > 0.8. We will also use TCGA datasets to identify the expression levels of selected optimal prognostic-relate genes of TNBC.

Construction and estimation of prognosis scoring model based on Hub Genes

We fit a multivariate Cox PHR model to the datasets with overall survival (OS) time to construct a lncRNA signature [15]. The best cut-off for TNBC samples stratification by this signature. Subsequently, the risk score coefficient was calculated based on the regression coefficient and expression level of prognostic lncRNA and was used as a survival risk measurement standard. The multivariate Cox regression model as following:

$$\text{Risk Score} = \sum i \cdot \text{coef}(\text{LncRNA}_i) \times \text{expr}(\text{LncRNA}_i)$$

Here, the coef represents the Cox PHR coefficient of lncRNA, and the expr represents lncRNA expression level.

We fit the lncRNA risk score model to the merged dataset of GSE42586 and GSE58812 with OS time to construct a lncRNA signature. According to the optimal cut-off value, TNBC samples in the data set were

classified into high-risk and low-risk groups. Recurrence-free survival was compared between the high- and low-risk groups by KM survival curve analysis and log-rank test with the Survival package in R [16]. The result was considered to figure a statistically significant difference in $P < 0.05$. Additionally, tROC analysis was executed to evaluate the predictive precision of multi-lncRNA prognostic signature and risk score.

Validation of the multi-lncRNA prognostic signature

To verify the predictive power and applicability of multi-lncRNA prognostic markers in TNBC, we collected TNBC-related TCGA ($n=96$) data sets as external validation. In the validation set, the coefficients of the three genes were used to calculate the risk score for each patient. Patients were divided into the high-risk group and low-risk group according to the median risk score of the training set. The prognostic characteristics of multiple lncRNAs were verified by KM survival analysis using a log-rank test and tROC analysis.

Functional annotation analysis

We carried out the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of significant modules by the R package clusterProfiler [17]. We did Gene Set Enrichment Analysis (GSEA) analysis of significant modules through the R packages clusterProfiler. To identify potential functions as significant modules, input annotation the complete Molecular Signatures Database (MSigDB) of mapping, the cut-off level of adjusted $P < 0.001$.

Results

Acquisition of BC expression datasets through array re-annotation

The detailed workflow is shown in Fig. 1. In this study, the 13 expression profiles data sets were included representing a total of 2950 samples of BC (we deleted 17 normal samples from GSE42586 and 25 normal or cell lines symbols from GSE65194) were obtained (Additional file 1: Table S1). There were two datasets available with survival information, including GSE42586 and GSE58812, with the OS. All data sets were normalized using RMA arithmetic.

A total of 23,520 microarray probes were matched using microarray annotation information (HG-U133A annotation). We are also getting 18115 re-annotation probes with gene biotypes. To stratify all 2950 BC samples into different molecular characteristics, including Basal-like (826 samples), Luminal A (904 samples), Luminal B (668 samples), Her2 (435 samples) and Normal-like (117 samples) intrinsic subtypes by PAM50 classifier. We collected 1217 patients with BC-related mRNA expression data and clinical data from the TCGA database. Based on the clinical data, the normal and TNBC groups were found to contain 113 and 96 samples, respectively.

Getting specific DEGs of TNBC

The other four, respectively, compared to normal-like subtypes, were selected and underwent DEGs analysis. Those candidate genes were set as background lists to get an intersection list of DEGs using Venn diagrams. The 3,468 DEGs with specific differences between Basal-like and Normal-like were obtained by removed the same gene probes come from the other three DEGs sets, which included 70 upregulated and 39 downregulated genes (Fig. 2a, Fig. 2b, Additional file 2: Table S2). The expression levels of all the genes were demonstrated, the top 100 DEGs (Head 50 and tail 50 genes) were rendered that were well clustered different subtypes of BC in the heatmap (Fig. 2c).

WGCNA co-expression network construction

Here, the weighted gene co-expression module was constructed from the 3484 DEGs, which included all TNBC subtypes samples of BC using the WGCNA algorithm. According to the scale-free topology criterion, we predicted a plot identifying scale-free topology in simulated expression data when the power value of $\beta = 16$ (scale-free $R^2 = 0.95$).

A total of 5 co-expression modules were identified (module size, ≥ 25 ; cut height, ≥ 0.99) by the average linkage hierarchical clustering based on TOM dissimilarity measure (1-TOM) stability, and every module was indicated in different colors (Fig. 3a). The four enriched modules (excluding grey modules that no genes had been assigned to any of the modules) have better connectivity (Fig. 3b). We delineate the eigengene network using a hierarchical cluster tree and a heatmap plot. A co-expression network of the associations between the subtype of BC and these modules was constructed (Fig. 3c). It is worth noting that we chose the blue module with the most significant correlation with TNBC as the relevant module and carried out the next analysis. There was a significant correlation between gene significance (GS) and module membership (MM), indicating that critical genes of the blue module were also highly associated with TNBC. (Fig. 3d).

Identification and validation of Hub Genes

We screened a total of 54 hub genes with high connectivity in the blue module based on absolute GS > 0.2, absolute MM > 0.8, including 5 lncRNA genes and 49 protein-coding genes (Fig. 4a). Among them, LINC00337, OSTM1-AS1, DEPDC1-AS1, DDX11-AS1, and LINC00608, had high functional significance in the hub genes. Additionally, based on the TCGA data, the expression levels of these four lncRNA genes were significantly higher in the TNBC tissues, including LINC00337, OSTM1-AS1, DEPDC1-AS1, and DDX11-AS1 ($p < 0.005$) (Fig. 4b, c, d, and e), especially the LINC00337, DEPDC1-AS1 and DDX11-AS1 ($p < 10^{-15}$). Surprisingly, there was no significant difference in the expression level of LINC00608 between healthy and TNBC tissues (Fig. 4f). The expression of LINC00337, DEPDC1-AS1, and DDX11-AS1 was

significantly up-regulated in the TNBC tissues ($p < 10^{-15}$). Therefore, the above three lncRNA were selected for further experimental verification.

Construction and estimation of prognosis scoring model based on 3 lncRNAs

We speculated that the identified complex 3-lncRNA signatures forcefully contribute to the prognosis of TNBC. Accordingly, we used this complex 3-lncRNA signature as an alone predictive mark to predict the risk of sample survival probability. We fit the lncRNA risk score model to the merged dataset of GSE42586 and GSE58812 with OS time to construct a lncRNA signature, the following formula: Risk Score = $0.2882 \times \text{expr}(\text{LINC00337}) - 1.1816 \times \text{expr}(\text{DEPDC1-AS1}) - 0.3307 \times \text{expr}(\text{DDX11-AS1})$.

To assess the robustness of the 3-lncRNA signature in predicting the risk of tumor recurrence for TNBC samples, the predictive power of the 3-lncRNA signature was then tested in the merged dataset. Next, the risk score of each patient was calculated, and patients were classified into a high-risk group ($n = 46$) and a low-risk group ($n = 57$) by the mid-cut point (Fig. 5a). The survival status of all patients and the heat maps of the three prognostic genes in the merged data set are shown in Fig. 5b and Fig. 5c, respectively. The KM survival curve showed that the high-risk group's OS was worse than that of the low-risk group (Fig. 5d). Besides, in the time-dependent ROC analysis, the prognostic characteristics of the three lncRNAs showed a more substantial area under the curve (AUC) value (Fig. 5e), indicating that the multi-lncRNA model had greater predictive ability in 1-year and 2-year OS.

Validation of the multi-lncRNA prognostic signature

To verify the predictive power and applicability of three-lncRNA prognostic markers in TNBC, we collected TNBC-related TCGA ($n = 96$) data sets as external validation. In the validation set, the three lncRNA coefficients were used to calculate the risk score for each patient. Consistent with the results in the training set, The KM curves of the external test sets were consistent with the results of the training set, and the high-risk group had a worse prognosis than the low-risk group (Fig. 6a). Time-dependent ROC analysis pointed out that AUC for 1-year, 2-year, and 5-year OS of the external validation set were 0.706, 0.828, and 0.731 (Fig. 6b). In summary, the prognostic characteristics of three-lncRNA showed an excellent performance in predicting OS in TNBC patients.

Functional annotation analysis

To further study the function of the blue modules containing three-lncRNA was highly correlated with TNBC, Gene Ontology (GO) term, KEGG pathway, and GSEA analysis were performed.

Biological processes analysis showed enrichment of GO terms associated with chromatin silencing at rDNA, protein heterotetramerization, interleukin-7-mediated signaling pathway and response to interleukin, etc. (Fig. 6c and Additional file 3: Table S3). According to KEGG pathway analysis, our results indicated that these genes were mainly involved in PD-L1 expression and PD-1 checkpoint pathway in cancer, Transcriptional misregulation in cancer, Cytosolic DNA-sensing pathway, and PI3K-Akt signaling pathway (Fig. 6d and Additional file 4: Table S4).

In the GSEA enrichment results (Fig. 7 and Additional file 5: Table S5), Results revealed that the genes of blue modules were enriched in transporter activity, Fischer dream targets, transmembrane transport and transmembrane transport activity. To sum up, these functional analysis results suggested that the blue module containing three-lncRNA was associated with the development and progress of TNBC.

Discussion

Breast cancer is the head cause of cancer death in women and easy to recrudescence. Among all the breast cancer subtypes, TNBC is more prevalent in young women, with large tumors, high lymphatic metastasis rate, and high clinical stage. Although BC treatment has developed during the last years, the knowledge to treat TNBC is still restricted due to lack expressed in ER, PR, and Her2. Although some studies have shown that some single biomarkers are strictly related to the occurrence and development of TNBC, it is complicated to predict the survival of patients with a single parameter. Therefore, establishing a comprehensive prognostic evaluation system, including multiple biomarkers, is the best way to improve prediction accuracy.

Although some BC related biomarkers have been proposed via WGCNA[18–20], it is rare to use WGCNA to further predict the potential biological targets of TNBC, especially in lncRNA. Tian et al. [21] only used co-expression analysis of lncRNAs with mRNAs to identify lncRNAs in TNBC. Dong et al. [22] identified the essential genes and pathways related to TNBC by DEGs and Protein-Protein Interaction (PPI) networks. These correlation network analyses only indicate whether multiple genes are related to each other. Moreover, the WGCNA can look for co-expressed gene modules and explore the correlation between gene networks and phenotypes of interest and the network's hub genes.

In this study, we tried to select more samples of breast cancer data from the GPL570 platform, a total of 2950 samples. Before the network analysis, we took 3484 genes with a significant difference between TNBC and normal-like candidate genes. After WGCNA analysis, we selected the blue module most related to TNBC, including 84 genes that were selected as hub genes with high connectivity. Additionally, based on the TCGA data, the expression levels of these three lncRNA genes were significantly higher in the TNBC tissues, including the LINC00337, DEPDC1-AS1, and DDX11-AS1 ($p < 10^{-15}$).

As far as we know, no study has used screening methods like ours to screen for three novel DEGs, all of which were positive prognostic lncRNAs. DDX11-AS1 (DDX11 Antisense RNA 1) is an RNA Gene that named "cohesion regulator non-coding RNA," or CONCR [23]. DDX11-AS1 affected the gene expressions

involved in HCC proliferation, differentiation, and cell cycle [24], which may be a novel oncogene in hepatocarcinogenesis by repressing LATS2 [25]. DDX11-AS1 is transcriptionally activated by MYC and is up-regulated in multiple cancer types, especially breast cancer and lung cancer [26]. The expression of DDX11-AS1 is cell cycle-regulated, and it is required for cell-cycle progression and DNA replication. There was a similar report in another article. The authors identified eight independent prognostic markers that independently predicted the survival of the samples in multiple cancers, including DDX11-AS1. These independent prognostic markers modulate cell cycle progression and cell proliferation [27]. Recent studies report that LINC00337 acts as the oncogene to promote gastric cancer cell proliferation through epigenetically repressing p21 mediated by EZH2 [28]. LINC00337 may up-regulate the expression of PBK and KIF23 through competitive binding of has-mir-373 and has-mir-519d. The competitive binding of has-mir-373 and has-mir-372 can up-regulate the expression of SLC7A11. The interaction between these RNAs may have an essential regulatory role in the immune infiltration in lung adenocarcinoma [29]. However, alterations of DEPDC1-AS1 in TNBC have been scarcely reported.

Moreover, a novel three-lncRNA-based prognostic model combining three-lncRNA signature was established and validated to improve survival prediction for TNBC patients. The patients in high-risk groups showed significantly poorer prognosis than the patients in the low-risk group. The multivariate Cox regression analysis showed that the three-lncRNA prognosis signature could be an independent factor in evaluating the prognosis. Internal and external validation was also conducted to confirm its predictive value. Further, Time-dependent ROC analysis of three-lncRNA showed favorable discrimination in both test set and validation set. Based on the GO and KEGG pathway analyses, the three-lncRNA may play crucial roles in chromatin silencing at rDNA, protein heterotetramerization, interleukin-7-mediated signaling pathway, PD-L1 expression and PD-1 checkpoint pathway in cancer and PI3K-Akt signaling pathway. The GSEA enrichment results revealed that the genes of blue modules were enriched in transporter activity, Fischer dream targets, transmembrane transport and transmembrane transport activity.

Conclusion

In summary, based on the comprehensive analysis of publicly available TNBC data in GEO and TCGA, we identified a novel, robust three-lncRNA-based prognostic model combining a three-lncRNA signature to predict 7-year OS in TNBC patients. Furthermore, the three-lncRNA signature can effectively identify low-risk patients from the high-risk group in TNBC patients. In other words, the three lncRNAs could be potential biomarkers in TNBC. The relevant lncRNA-based model could predict the specific survival rate and promote the choice of specific treatment selections of TNBC patients.

Abbreviations

TNBC: Triple-negative breast cancer

BC: Breast cancer

GEO: Gene Expression Omnibus

WGCNA: Weighted correlation network analysis

TCGA: The Cancer Genome Atlas

ROC: Receiver operating characteristic curve

KM: Kaplan-Meier

lncRNA: Long non-coding RNA

ER: Estrogen receptor

PR: Progesterone receptor

Her2: Human epidermal growth factor receptor 2

TNM: Tumor node metastasis

OS: Overall survival

DEGs: Differentially expressed genes

MM: Module membership

GS: Gene significance

GO: Gene Ontology

KEGG: Kyoto Encyclopedia of Genes and Genomes

GSEA: Gene Set Enrichment Analysis

MSigDB: Molecular Signatures Database

Declarations

Authors' contributions

CY, LC, designed the experiments and interpreted the data. CY, HY, and MS conducted bioinformatics and statistical analyses. CY and WT wrote the manuscript. All authors have read and approved the manuscript for publication.

Funding

No funding was received.

Availability of data and material

The datasets generated during and/or analyzed during the current study are available in the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>) and TCGA database (<https://portal.gdc.cancer.gov/>).

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Acknowledgements

The information of this study here is obtained by the Cancer Genome Atlas database (TCGA) and Gene Expression Omnibus (GEO). We are grateful to them for the source of data used in our study.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin.* 2011;61(2):69-90. doi:10.3322/caac.20107.
2. Navratil J, Fabian P, Palacova M, Petrakova K, Vyzula R, Svoboda M. [Triple Negative Breast Cancer]. *Klin Onkol.* 2015;28(6):405-15.
3. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* 2011;121(7):2750-67. doi:10.1172/JCI45014.
4. Zhang Z, Lin E, Zhuang H, Xie L, Feng X, Liu J et al. Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma. *Cancer Cell Int.* 2020;20:27. doi:10.1186/s12935-020-1113-6.

5. Gettman MT, Blute ML, Spotts B, Bryant SC, Zincke H. Pathologic staging of renal cell carcinoma: significance of tumor classification with the 1997 TNM staging system. *Cancer*. 2001;91(2):354-61. doi:10.1002/1097-0142(20010115)91:2<354::aid-cncr1009>3.0.co;2-9.
6. Moreira MP, da Conceicao Braga L, Cassali GD, Silva LM. STAT3 as a promising chemoresistance biomarker associated with the CD44(+/-high)/CD24(-/low)/ALDH(+) BCSCs-like subset of the triple-negative breast cancer (TNBC) cell line. *Exp Cell Res*. 2018;363(2):283-90. doi:10.1016/j.yexcr.2018.01.018.
7. Li HY, Liang JL, Kuo YL, Lee HH, Calkins MJ, Chang HT et al. miR-105/93-3p promotes chemoresistance and circulating miR-105/93-3p acts as a diagnostic biomarker for triple negative breast cancer. *Breast Cancer Res*. 2017;19(1):133. doi:10.1186/s13058-017-0918-2.
8. Pedersen MH, Hood BL, Ehmsen S, Beck HC, Conrads TP, Bak M et al. CYPOR is a novel and independent prognostic biomarker of recurrence-free survival in triple-negative breast cancer patients. *Int J Cancer*. 2019;144(3):631-40. doi:10.1002/ijc.31798.
9. Prat A, Adamo B, Cheang MC, Anders CK, Carey LA, Perou CM. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *Oncologist*. 2013;18(2):123-33. doi:10.1634/theoncologist.2012-0397.
10. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71. doi:10.1093/nar/gkv1507.
11. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-7. doi:10.1200/JCO.2008.18.1370.
12. Gendoo DM, Ratanasirigulchai N, Schroder MS, Pare L, Parker JS, Prat A et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2016;32(7):1097-9. doi:10.1093/bioinformatics/btv693.
13. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007.
14. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559. doi:10.1186/1471-2105-9-559.
15. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385-95. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.
16. Schroder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*. 2011;27(22):3206-8. doi:10.1093/bioinformatics/btr511.
17. Xiong Z, Ye L, Zhenyu H, Li F, Xiong Y, Lin C et al. ANP32E induces tumorigenesis of triple-negative breast cancer cells by upregulating E2F1. *Mol Oncol*. 2018;12(6):896-912. doi:10.1002/1878-0261.12202.

18. Shi H, Zhang L, Qu Y, Hou L, Wang L, Zheng MJOL. Prognostic genes of breast cancer revealed by gene co-expression network analysis. 2017;14(4):4535-42.
19. Tang J, Kong D, Cui Q, Wang K, Zhang D, Gong Y et al. Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis. *Front Oncol*. 2018;8:374. doi:10.3389/fonc.2018.00374.
20. Li J, Wang W, Xia P, Wan L, Zhang L, Yu L et al. Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in breast cancer patients. 2018.
21. Tian T, Gong Z, Wang M, Hao R, Lin S, Liu K et al. Identification of long non-coding RNA signatures in triple-negative breast cancer. *Cancer Cell Int*. 2018;18:103. doi:10.1186/s12935-018-0598-8.
22. Dong P, Yu B, Pan L, Tian X, Liu F. Identification of Key Genes and Pathways in Triple-Negative Breast Cancer by Integrated Bioinformatics Analysis. *Biomed Res Int*. 2018;2018:2760918. doi:10.1155/2018/2760918.
23. Sabatier R, Finetti P, Adelaide J, Guille A, Borg JP, Chaffanet M et al. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. *PLoS One*. 2011;6(11):e27656. doi:10.1371/journal.pone.0027656.
24. Marchese FP, Grossi E, Marin-Bejar O, Bharti SK, Raimondi I, Gonzalez J et al. A Long Noncoding RNA Regulates Sister Chromatid Cohesion. *Mol Cell*. 2016;63(3):397-407. doi:10.1016/j.molcel.2016.06.031.
25. Liao HT, Huang JW, Lan T, Wang JJ, Zhu B, Yuan KF et al. Identification of The Aberrantly Expressed LncRNAs in Hepatocellular Carcinoma: A Bioinformatics Analysis Based on RNA-sequencing. *Sci Rep*. 2018;8(1):5395. doi:10.1038/s41598-018-23647-1.
26. Li Y, Zhuang W, Huang M, Li X. Long noncoding RNA DDX11-AS1 epigenetically represses LATS2 by interacting with EZH2 and DNMT1 in hepatocellular carcinoma. *Biochem Biophys Res Commun*. 2019;514(4):1051-7. doi:10.1016/j.bbrc.2019.05.042.
27. Ali MM, Akhade VS, Kosalai ST, Subhash S, Statello L, Meryet-Figuere M et al. PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. *Nat Commun*. 2018;9(1):883. doi:10.1038/s41467-018-03265-1.
28. Hu B, Wang X, Li L. Long noncoding RNA LINC00337 promote gastric cancer proliferation through repressing p21 mediated by EZH2. *Am J Transl Res*. 2019;11(5):3238-45.
29. Wei B, Kong W, Mou X, Wang S. Comprehensive analysis of tumor immune infiltration associated with endogenous competitive RNA networks in lung adenocarcinoma. *Pathol Res Pract*. 2019;215(1):159-70. doi:10.1016/j.prp.2018.10.032.

Figures

Figure 1

The workflow of the selection process for TNBC survival-related 3-lncRNA signature.

Figure 2

(a) Venn diagram showing specific genes targeted by the panel of differentially DEGs in molecular subtypes of BC. (b) Volcano plot of 3484 genes. Red and blue plots respectively represented genes with $\logFC > |1.0|$ in the $p < 0.01$. Yellow plots displayed the rest of the genes with no significant expression change. (c) Heatmap of the top 100 DEGs (head 50 and tail 50 genes). Red and green respectively displayed high and low expression.

3

Figure 3

Construction of co-expression modules of TNBC related genes. (a) The construction of co-expression modules by R software. Each branch in the figure displayed one gene, and each color below represented one co-expression module. (b) interaction analysis of co-expression genes. Different colors of abscissa and ordinate represented different modules. (c) Heatmap of the correlation between module eigengenes (MEs) and a subtype of breast cancer. Each cell reports the correlation (and p-value) between the module eigengenes (rows) and subtype (columns). (d) A scatterplot of gene significance for the weight (GS) versus module membership (MM) in the blue module.

Figure 4

Identification and validation of Hub Genes. (a) The 54 hub-genes among the whole network. The red nodes represented lncRNA genes, and yellow nodes represented protein-coding genes. Expression levels of differential expressed lncRNAs in different groups of TCGA datasets: (b) LINC00337, (c) OSTM1-AS1, (d) DEPDC1-AS1, (e) DDX11-AS1, and (f) LINC00608.

Figure 5

Prognostic analysis of three-lncRNA signature in the training set. The dotted line represented the median risk score and divided the patients into low- and high-risk group. (a) The curve of risk score. (b) Survival status of the patients. More dead patients corresponding to the higher risk score. (c) Heatmap of the expression profiles of the five prognostic genes in low- and high-risk group. (d) Kaplan–Meier survival analysis of the five-gene signature. (e) Time-dependent ROC analysis the of the five-gene signature.

Figure 6

Validation of the three-lncRNA signature. (a) TCGA(n=96) was regarded as the external validation set. Kaplan-Meier survival analysis of the three-lncRNA signature in external validation set. (b) Time-dependent ROC analysis of the three-lncRNA signature in external validation set. GO and KEGG analysis of genes from the blue modules. (c) Scatter plot depicting the relationship between genes and GO terms, the enrichment results of the top 20 are visualized. (d) overlapping hierarchical links plot indicates the relationship between genes and KEGG pathways; the enrichment results of the top 23 are visualized.

Figure 7

GSEA enrichment results of genes from the blue modules ($p < 0.05$). GSEA validated enhanced activity of (a) "GO_TRANSPORTER_ACTIVITY", (b) "FISCHER_DREAM_TARGETS", (c) "GO_TRANSMEMBRANE_TRANSPORT", (d) "GO_TRANSMEMBRANE_TRANSPORT_ACTIVITY".

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile5.xlsx](#)
- [Additionalfile4.xlsx](#)
- [Additionalfile3.xlsx](#)
- [Additionalfile2.xlsx](#)
- [Additionalfile1.xlsx](#)