Iran. J. Biotechnol. July 2024;22(3): e3853 DOI: 10.30498/ijb.2024.432171.3853

Research Article

Evaluation of lncRNAs as Potential Biomarkers for Diagnosis of Metastatic Triple-Negative Breast Cancer through Bioinformatics and Machine Learning

Shiva Soleimani[1], Farkhondeh Pouresmaeili,[2,3,*] Iman Salahshoori Far[1]

[1]Department of Biology, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Men's Health and Reproductive Health Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Medical Genetics Department, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[*]*Corresponding author*: Farkhondeh Pouresmaeili, Men's Health and Reproductive Health Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran; Medical Genetics Department, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran; Tel/Fax: +98-2123872572, E-mail: Pouresmaeili_g@sbmu.ac.ir

Received: 2023/12/24 ; **Accepted:** 2024/07/20

Background: Triple-negative breast cancer (TNBC) is highly invasive and metastatic to the lymph nodes. Therefore, it is an urgent priority to distinguish novel biomarkers and molecular mechanisms of lymph node metastasis as the first step to the disease investigation. Long non-coding RNAs (lncRNAs) have widely been explored in cancer tumorigenesis, progression, and invasion.

Objectives: This study aimed to identify and evaluate lncRNAs in the signaling pathway of *MMP11* gene in both metastatic and non-metastatic TNBC samples. The potential of lncRNAs in prognosis and diagnosis of the disease was also assessed using bioinformatics analysis, machine learning, and quantitative real-time PCR.

Materials and Methods: Using machine learning algorithms, we analyzed the available BC data from the Cancer Genome Atlas Network (TCGA) and identified three potential lncRNAs, gastric adenocarcinoma-associated, positive CD44 regulator, long intergenic noncoding RNA (*GAPLINC*), *TPT1-AS1*, and *EIF1B* antisense RNA 1 (*EIF1B-AS1*) that could successfully distinguish between metastatic and non-metastatic TNBC.

Results: The results showed the upregulation of *GAPLINC* lncRNA in metastatic BC tissues, compared to non-metastatic ($P<0.01$) and normal samples, though *TPT1-AS1* and *EIF1B-AS1* were downregulated in metastatic TNBC samples ($P<0.01$).

Conclusion: Given the aberrant expression of candidate lncRNAs and the underlying mechanisms, the above-mentioned RNAs could act as novel diagnostic and prognostic biomarkers in metastatic BC.

Keywords: Biomarkers, Breast neoplasms, Long noncoding RNA, Neoplasm metastasis, Triple negative breast neoplasms

1. Background

Breast cancer (BC) is the most commonly diagnosed cancer and the leading cause of cancer malignancy among females worldwide. The BC burden is still increasing in form of incidence and mortality and continues to have a large impact on the global number of cancer deaths (1). BC is categorized into three main subtypes based on the presence or absence of proteins in BC cells. Hormone receptor-positive BC, which has either estrogen receptor or progesterone receptor protein in cancer cells, includes 70% of BC patients. ERBB2-positive or HER2-positive BC has high levels of ERBB2 protein in cancer cells and constitutes for 15%-20% of BC cases. Triple-negative breast cancer

(TNBC) entails 15% of BC cases and does not have estrogen, progesterone, or ERBB2 protein on the surface of cancer cells (2). Among all the BC subtypes, TNBC is more complex and accounts for a major cause of gynecological cancer deaths, associated with more positive lymph nodes, higher tumor grading, and poorer prognosis (3). A vast number of patients with TNBC relapse rapidly and often develop visceral (including liver, lung, and brain) metastasis (4).

Current targeted options for treating BC include hormonotherapy and immunotherapy, as well as PARP inhibitors; however, there is not such therapy for TNBC, mainly due to the lack of predictive biomarkers (5). Recently, remarkable advances have been made toward the early diagnosis and intervention, i.e. the identification of *MMP-11* gene as a metastatic BC biomarker that is overexpressed in cancer cells, stromal cells, and the adjacent microenvironment. *MMP-11* belongs to the matrix metalloproteinase family (MMPs), known for their overexpression in cancer cells, stromal cells, and the surrounding microenvironment. MMPs are zinc-dependent endopeptidases responsible for degrading the extracellular matrix, thus aiding in the breakdown of the basal membrane and connective tissue matrix. This activity is crucial in cancer progression and metastasis. *MMP-11* expression has been found to be elevated and more varied in metastatic specimens compared to non-metastatic tumor samples. Moreover, diagnostic kits for cancer metastasis are developed based on the expression level of this gene and are currently in clinical trials (6-8). Despite this progress, the study of novel candidate genes involved in TNBC progression and prognosis and the role of molecular mechanisms in tumorigenesis need further illumination. Moreover, for the improvement of the survival of TNBC patients, identification of predictive biomarkers is essential, in order to evaluate the risk of metastasis, assess response to therapies and develop new therapeutic methods (9).

Recently, investigations across the species have demonstrated that eukaryotic genomes transcribe a wide range of RNAs, including long noncoding RNAs (lncRNAs), protein coding mRNAs, and short non-coding transcripts. LncRNAs are recognized as RNA molecules comprising of transcripts with more than 200 nucleotides in length, though they lack the ability of coding proteins. These molecules have also been experimentally characterized to show their distinct cellular functions (10). LncRNAs have been suggested as significant regulators for promoting or preventing tumor progress and play a crucial regulatory role in terms of transcriptional, post-transcriptional, and epigenetic levels. Therefore, mutations or aberrant expression of lncRNAs have been correlated with various malignant biological processes, comprising carcinogenesis, cell proliferation, migration, invasion, and apoptosis (4). Evidence of novel and potentially beneficial biomarkers has revealed that the abnormal expression of these RNAs is closely related to TNBC development and invasion (11).

2. Objective

This study attempted to explore candidate lncRNAs that correlate with *MMP11* signaling pathways and metastatic BC in TNBC samples, using a comprehensive bioinformatics analysis, machine learning, and quantitative real-time PCR. For the first time in the present research, the expression level and the role of lncRNA *EIF1B-AS1* biomarker have been investigated on BC samples.

3. Materials and Methods

3.1. Data Collection and Preparation

The Cancer Genome Atlas (TCGA, https://www.cancer.gov/) (12, 13) and Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) (14, 15) were the datasets used in the current study. From the TCGA-BRCA dataset, the RNA-Seq (raw HTseq) data (16)with corresponding clinical data were downloaded from the TCGABiolinks R package (15, 17). Also, five microarray datasets, including GSE20685, GSE19615, GSE17907, GSE16446, and GSE6532, were selected from GEO databases (15, 18) and generated from GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) platform (14, 15, 19). For RNA-seq data, we used the edgeR package to eliminate the genes with zero or near-zero expression, based on the count-per-million criterion, which was less than 10% in 70% of the samples (20, 21). The normalization of the data was then carried out using the TMM method, and finally, the data were transferred to the logarithmic mode (22, 23). For further analysis, the normalized expression matrix was utilized. Also, microarray data (CEL files) of each dataset were downloaded. Subsequently, using the affy and limma packages, data reading, data transfer to logarithmic mode, and data normalization using the

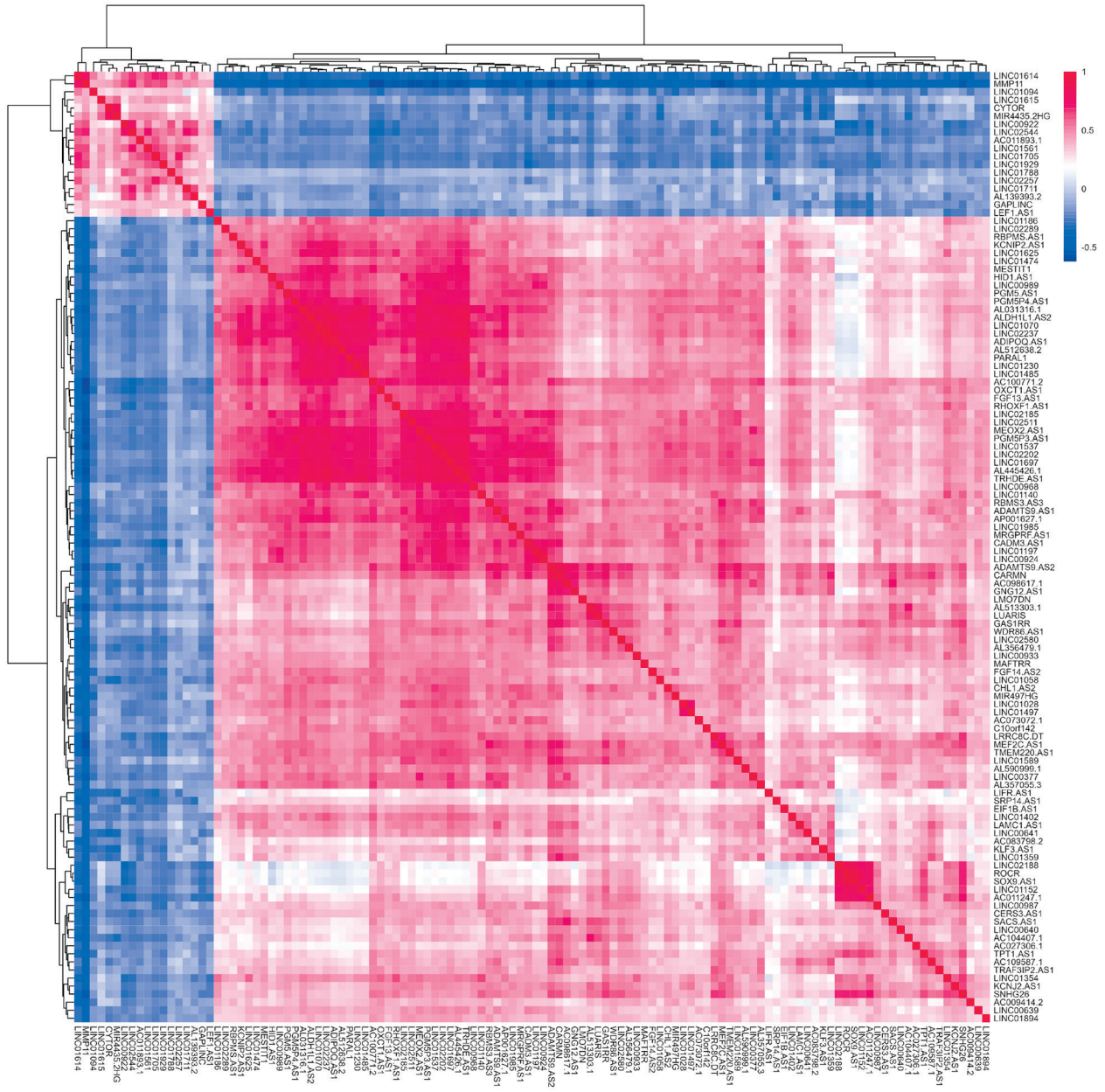

Figure 1. The heatmap of lncRNAs genes. The genes (n = 117) were grouped into two main clusters using ggplot2 R package.

RMA method, background correction was performed (24). In the end, all the selected data from each dataset were merged with each other, and the batch effects were removed using the Combat function provided by the SVA package version 3.48.0 (25). The obtained expression matrix was employed for all the analyses.

3.2. Differential Gene Expression Analysis

TCGA samples were classified into two (normal and patient) groups based on their clinical data. The total number of samples was 1,222, i.e. 113 samples with normal and 1109 samples with cancer tissues. We applied edgeR and limma packages in the R environment for differential gene expression analysis to fit the linear model on the expression data (26, 27). Only genes that remained from the count-per-million gene exclusion filter were considered for differential gene expression analysis. The lncRNA expression

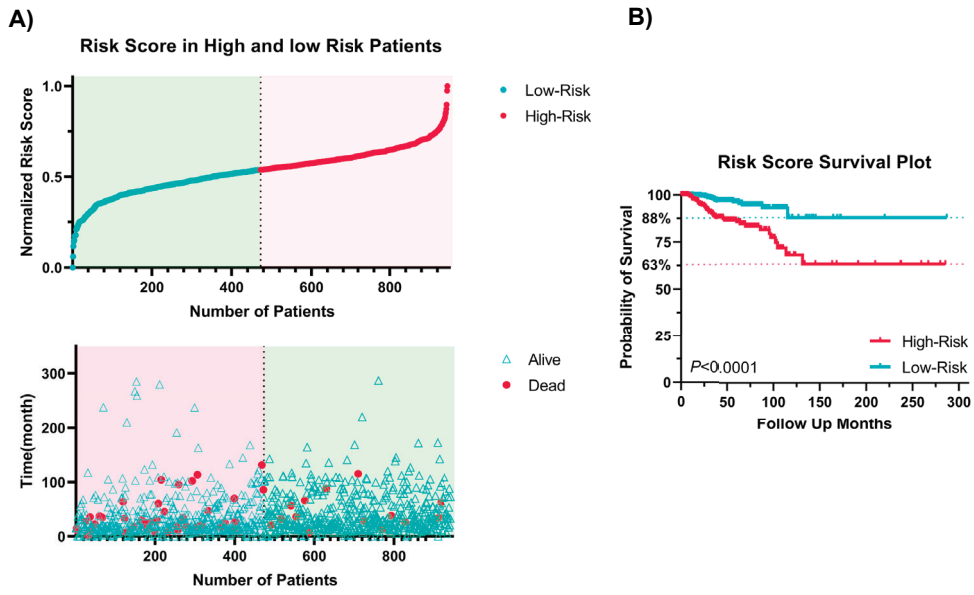

Figure 2. Cox regression analysis. Multivariate Cox regression model in patients' survival rate considering two crucial factors, the pathological stage and age, indicating risk scores **A)** in high- and low-risk patients and **B)** of the survival plot.

matrix was extracted from the normalized expression matrix, and then the spearman correlation test was performed between all lncRNAs and *MMP11*, a very significant gene related to BC metastasis.

3.3. Receiver Operating Characteristic (ROC) Curve and Diagnosis Analysis

The pROC package in R was used to draw ROC curves. For this purpose, the genes correlated with *MMP11* were classified into two main clusters through the K-means hierarchical clustering method. The second cluster that contained more genes was divided into nine subclusters in order to have higher ROC curves. Finally, two tables were created, the first included the labels of samples (1: patient and 0: healthy), and the second included the transposed expression matrix on 117 lncRNAs plus *MMP11* (columns show corresponding genes, and rows are samples). The sensitivities and specificities of the classification model were calculated in the pROC package, and ROC curves were generated using the ggplot2 R package.

3.4. Evaluation of Candidate LncRNAs as Biomarkers for BC Diagnosis and Progression

To evaluate the impact of candidate lncRNAs on BC progression and their potential application as diagnosis markers, we first performed a ROC curve analysis. Then the 117 genes were grouped into two main clusters, as shown in the heatmap (**Fig. 1**). The first cluster contained 18 genes, including *MMP11* and its positively correlated genes; however, the second cluster included 100 genes, most of which were negatively correlated with *MMP11*. We executed ROC analysis on each cluster separately, and to depict it better, we subdivided the second cluster into nine smaller segments, each containing a different number of genes.

3.5. Survival Analysis of Samples

To perform the survival analysis of samples, we eliminated all 113 normal samples from the gene expression dataset, and only 1109 patients remained. The patients' clinical data were added to the expression matrix. Patients with unknown conditions and those with non-BC deaths were removed from further investigation. Overall, 945 patients remained that were divided into two groups (dead: 58 and alive: 887). Using Cox regression model, we analyzed the survival of the patients. First, the univariate Cox proportional hazard model was performed for all genes to find survival-related candidate lncRNAs (28, 29). The expression

level of each lncRNA was scaled between zero and one, separately, with the following formula (30):

Where "X" indicates the expression of a specific lncRNA, "i" stands for patients' number, and min and max show minimum and maximum levels of lncRNAs expression in all the patients. Moreover, the risk score was measured based on the beta value related to survival-related lncRNAs using the following formula (31):

Where "$W_{j"}$ implies the univariate coefficient for lncRNA j, "$exp_{ij"}$ indicates the scaled (between 0 and 1) expression value of lncRNA j in patient i, and "n" denotes the number of testings lncRNAs. To evaluate the independence of the risk score model, we applied a multivariate Cox regression model for patients' survival rate considering two crucial factors, the pathological stage and age. In addition, using the median of risk scores as the cut-off value, we classified the patients into high-risk and low-risk groups (**Fig. 2A**). The univariate test allowed us to find significant genes related to the survival prognosis of BC patients. The criteria for the selection of these genes were LRT results, which revealed that 15 out of 117 lncRNAs were significantly associated with patients' survival (**Table 1**).

3.6. Deep Learning Model Training

A new expression matrix was generated from five microarray datasets containing data from patients with non-metastatic and metastatic breast tumors. The matrix included a total of 700, 457 non-metastatic and 243 metastatic, samples. TCGA data were transferred to logarithmic scale using the following formula: ; where X stands for scaled expression for each gene, and C stands for raw expression value. We considered only 51 lncRNAs in our matrix because other lncRNAs did not exist in the GPL570 platform. By applying Keras and Tensorflow libraries in the python environment, we used both datasets to train deep learning models for classification of patients into non-metastatic and metastatic cases. Ultimately, four deep learning models were generated using the mentioned libraries. We compared the performance of all the models, and one model with the highest capability was selected. We considered 15% of our total data for testing, 15% for validating and 70% for training datasets. The model contained four hidden layers; the first layer consisted of 51 units, and the second, third, and four layers included 40, 25, and 10 units, respectively. The activation function for all the hidden layers was tanh, and the output layer was sigmoid. RMSprop was selected as the

Table 1. Results of Cox regression and unvariant test.

Ensembl Ids	Beta	LRT	HR	Hrlower	HRupper	Name
ENSG00000241158	-0.3824	0.0183	0.6822	0.4765	0.9766	ADAMTS9-AS1
ENSG00000225670	-0.3261	0.0284	0.7216	0.5261	0.9900	CADM3-AS1
ENSG00000222041	0.3753	0.0049	1.4554	1.1220	1.8879	CYTOR
ENSG00000226031	-0.3280	0.0408	0.7203	0.4967	1.0446	FGF13-AS1
ENSG00000232284	-0.3383	0.0215	0.7129	0.5268	0.9648	GNG12-AS1
ENSG00000179428	-0.4284	0.0140	0.6515	0.4371	0.9711	IL6-AS1
ENSG00000254862	-0.3375	0.0415	0.7134	0.4928	1.0329	LGR4-AS1
ENSG00000244968	-0.3634	0.0096	0.6953	0.5241	0.9223	LIFR-AS1
ENSG00000237248	-0.4484	0.0033	0.6386	0.4569	0.8925	LINC00987
ENSG00000225039	-0.2927	0.0401	0.7462	0.5531	1.0066	LINC01058
ENSG00000223485	0.3837	0.0010	1.4678	1.1996	1.7960	LINC01615
ENSG00000232679	0.3596	0.0020	1.4328	1.1605	1.7690	LINC01705
ENSG00000267013	0.2554	0.0384	1.2910	1.0283	1.6208	LINC01929
ENSG00000229108	-0.3791	0.0359	0.6844	0.4561	1.0270	LINC02587
ENSG00000172965	0.4961	0.0001	1.6423	1.2750	2.1154	MIR4435-2HG

optimizer, and the learning rate of 0.00003 was chosen to determine the speed of model training. As our model performed only binary classification, we used binary cross entropy as loss function and accuracy as reporting metrics. The model was trained for 3000 epochs, and the bias was considered for all layers at zero at the beginning of training. We used 19% of the remaining training data to fit the model as the validation dataset. Therefore, the model was first trained based on the training dataset, then compared to validation data and finally double-checked with the test dataset. In the end, a model with high accuracy and low loss was selected. To avoid overfitting, we used dropout layers with 0.3 rate. Thus, we ensured that the model will focus on general features, instead of attending to only certain features.

3.7. Sample Collection

Overall, 60 BC samples and 10 normal breast samples were acquired from the Iranian Tumor Bank, Cancer Institute of Iran, Tehran. The samples did not receive any drugs or therapy, and all cancer samples were confirmed by a pathologist and stored in liquid nitrogen until use. The samples' clinical information is illustrated in **Table 2**.

3.8. RNA Extraction, Complementary DNA Synthesis, and RT-Quantitative PCR

RNA was isolated from each sample using Trizol (Invitrogen, Germany) in conformity with the manufacturer's instructions after three washes with PBS for the removal of contaminants and necrotic cells. Based on the procedure of Invitrogen DNase kit, DNA was excluded from the extracts. Afterwards, the TaKaRa kit (USA) was used to synthesize complementary DNA. Specific primers, as shown in the **Supplementary Table 1**, were designed by Primer-BLAST tool (www.ncbi.nlm.nih.gov/tools/primer-blast). The expression of the mentioned genes (*GAPINC*, *EIF1B-AS1*, *TPT1-AS1*, *and MMP11*) in normal and cancer samples was quantified by the aid of RT-quantitative PCR with specific primers and SYBR Green master mix. Beta-actin was employed as an internal reference, and $2^{-\Delta Ct}$ was utilized to determine the gene expression in each sample.

3.9. Statistics and Software

All the data preprocessing and analysis were accomplished by the R programming language (version 4.0.2). GhraphPad v8 was employed to draw and display charts. Differential expression analysis was performed using the Limma package in R. Normalized gene expression data were inputted, and a design matrix was specified to account for experimental conditions. Limma fitted a linear model to the data, incorporating empirical Bayes moderation to stabilize variance estimates. Hypothesis testing was conducted to identify differentially expressed genes, with multiple testing correction applied to control false discovery rate. Significant genes were determined based on adjusted p-values. False discovery rate level considered in all analysis was less than 0.01. The log-rank test (LRT) was also employed to evaluate the significance of candidate gene expression, which was considered as LRT threshold of 0.05, in patients' prognosis.

4. Results

4.1. Identification of Significant lncRNAs in BC Metastasis

The TCGA RNA-seq dataset was applied to find significant lncRNAs that may be important in BC metastasis. According to TCGA RNA-seq, the expression matrix contained 56,602 genes, among which 29,833 were removed because of low expression levels in breast tissue. Thus, only 267,69 genes were kept for our study (**Fig. 3A**). In both cancer and normal

Table 2. Results of multivariate Cox regression

Factor	Beta	LRT	HR lower	HR	HR upper
TS	1.3256	<0.0001	1.9698	3.7645	7.1942
Risk score	2.1334	<0.0001	3.4898	8.4472	20.1284
Age	0.5064	0.0561	0.9672	1.6502	2.8612
Pathological stage	1.1941	<0.0001	2.2465	4.8533	3.3129

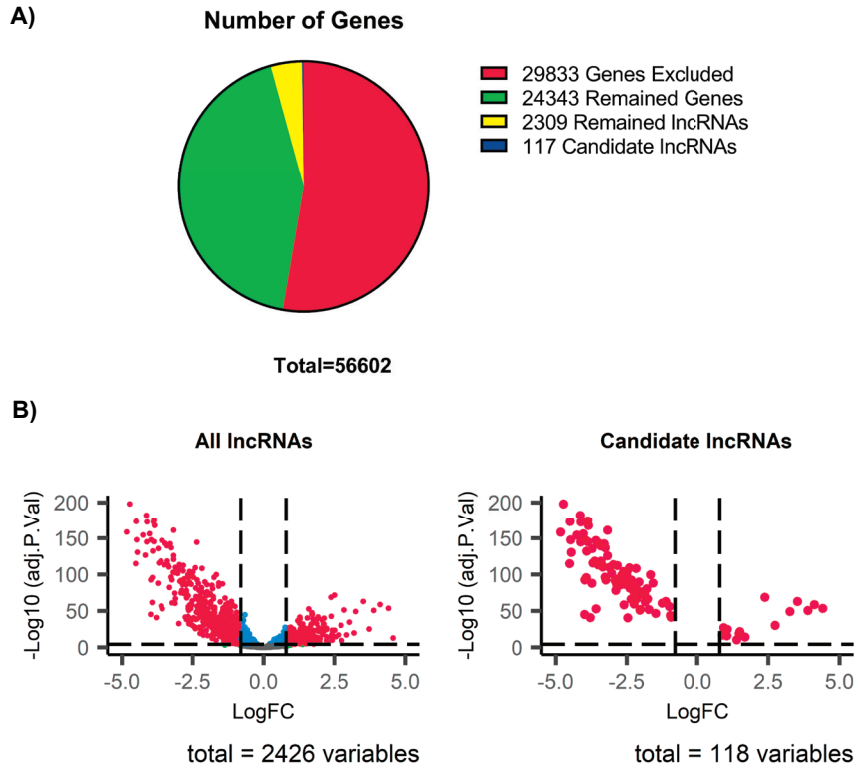

Figure 3. TCGA RNA-seq dataset results. **A**) Results of TCGA RNA-seq and expression matrix to identify the number of genes related to breast cancer; **B**) Analysis of lncRNAs expression in cancer and normal samples, showing that 117 lncRNAs were significantly correlated with *MMP11* expression level.

samples, we found that 117 lncRNAs were significantly correlated with *MMP11* expression level (correlation coefficient > 0.4 or <-0.4; P<0.01). The list of 117 candidate lncRNAs is shown in **supplementary Table 2**. To find significantly deregulated genes between the normal and patients groups, selected based on their corresponding clinical data, we analyzed a differential gene expression, and the results revealed that all the 117 candidate lncRNAs and *MMP11* were considerably up- or downregulated, when compared to the normal and cancer groups (adj.*P*.Val < 0.0001 & LogFC [log fold change] > |0.8|; **Fig. 3B**).

4.2. ROC Curve and Diagnosis Analysis

The results of ROC curve revealed that most of our candidate lncRNAs are potential biomarkers for BC diagnosis. The area under curves (AUCs) indicated five lncRNAs as fair (0.7<AUC<0.8), 19 as good (0.8<AUC<0.9), and 92 as excellent potential biomarkers (0.9<AUC<1). The only poor biomarker with the lowest AUC (0.697) was *LINC01788*, and the excellent biomarkers, except for *MMP11*, were *LINC01614*, *TMEM220.AS1*, *ADAMTS9.AS2*, and *LINC02511* with AUC>0.98.

4.3. Prediction of Patients' Survival Rate by LncRNAs

A multivariate Cox regression analysis was performed to show the dependency of the calculated risk scores on other important clinical characteristics. According to Cox regression and univariant test, 10 genes showed good prognosis behavior, whereas five genes had poor prognosis (**Table 1**). In addition, a risk score model was computed based on survival-related lncRNAs. In order to avoid the impact of high and low expression levels of lncRNAs, we scaled the expression level of lncRNAs between zero and one. The outcomes indicated that the risk score model could dependently predict the patients' survival rate, and the combination of 15 lncRNAs

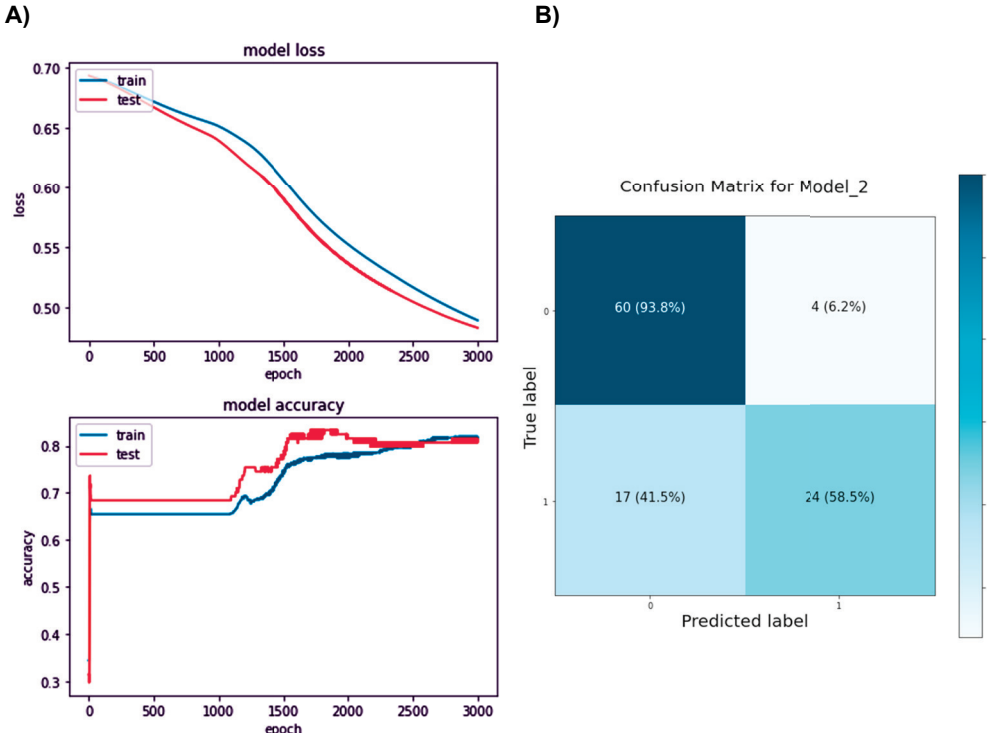

Figure 4. Deep learning analysis. **A**) Model accuracy and loss; **B**) confusion matrix using creat R package.

expression levels is related to their survival state (**Table 1**). In the Cox regression analysis, the cut-off value was considered as median risk scores, and based on this criterion, the patients were categorized into two (high-risk and low-risk) groups (**Fig. 2A**). A survival plot was drawn to distinguish the survival rates of patient groups (**Fig. 2B**). Results revealed that 63% (44 out of 58 deaths) and 88% (only 14 deaths) of the patients were in the high- and low-risk group, respectively. The results also suggested that 15 out of 117 lncRNAs could be considered as prognostic biomarkers and could have a role in the BC development and malignancy.

4.4. A Deep Learning Model with High Capability for Classification of Patients

The model was trained several times and finally showed satisfactory results, with 82% training accuracy, 81% validation accuracy, and 70% test accuracy. To avoid overfitting, we simultaneously checked the validation and training data loss and accuracy (**Fig. 4A**). We also generated the confusion matrix (**Fig. 4B**) and computed other parameters to assess our model with details. Calculations suggested that the weighted F1 score of this model is 0.79, which is highly acceptable for classification of patients. Finally, those lncRNAs that had the most significant role in the model, in both methods, were evaluated. According to the decision tree, six lncRNAs (*TPT1-AS1*, PGM5.AS1, R4435.2HG, FIF1R.AS1, *GAPLINC*, and LUARIS) indicated the most important role in the model and in separation of metastatic from non-metastatic samples (**Fig. 5A).** On the other hand, the results of support vector machine showed that *TPT1-AS1*, LINC01099, *GAPLINC*, LINC01140, and *EIF1B-AS1* had the most essential function in the model (**Fig. 5B**). These findings suggest that the mentioned lncRNAs could have a crucial role in the metastasis of BC.

4.5. Significant Expression Changes of GAPLINC, TPT1-AS1, and EIF1B-AS1 in Samples

Since *GAPLINC, TPT1-AS1,* and *EIF1B-AS1* have been less studied in BC, they were taken into consideration in this study. In this regard, MMP11, *GAPLINC*, *TPT1-AS1*, and *EIF1B-AS1* expression levels were investigated in metastatic cancer samples compared to

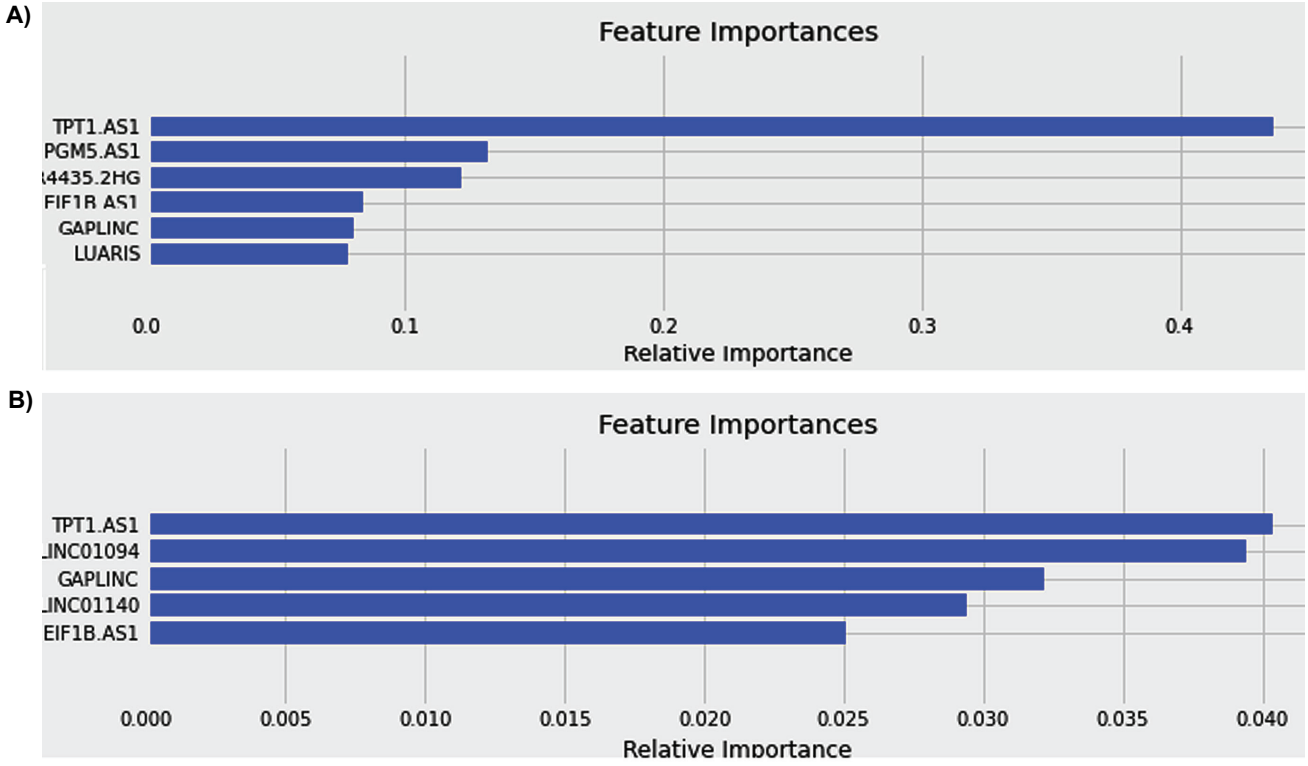

Figure 5. **Decision tree constructed by R packege rpart. A**) The genes that play the most important role in the model and in separation of metastatic from non-metastatic samples; **B**) The results of SVM showed that *TPT1-AS1*, LINC01099, *GAPLINC*, LINC01140 and *EIF1B-AS1* had the most essential role in the model.

normal and non-metastatic BC samples. Data showed that the expression level of *MMP11* and *GAPLINC* in cancer, but not normal, samples increased significantly (**Fig. 6**; FDR<0.01; logFC>0.8). This result was similar to our ex vivo studies (**Fig. 6**; P<0.01). Our results also revealed that the expression level of MMP11 and *GAPLINC* increased in metastatic compared to non-metastatic samples (**Fig. 6A and 6B**; P<0.01). However, the expression level of *TPT1-AS1* and *EIF1B-AS1* decreased in both TCGA and ex vivo data and in metastatic relative to non-metastatic samples (**Fig. 6C and 6D**; P<0.01). These data display that *GAPLINC*, *TPT1-AS1*, and *EIF1B-AS1* could have a role in BC malignancy through the process of metastasis.

5. Discussion

TNBC, in comparison to other BCs, is characterized by high malignancy, easy recurrence, young onset, and low survival rates. While initial response to chemotherapy might be more profound, relapse in visceral organs, including the central nervous system, is common, though its underlying mechanism has not been revealed properly (32, 33). Therefore, targeting lncRNAs could be an opportune clinical approach in metastatic TNBC diagnosis and treatment, especially in early stages.

Recent studies have shown the potential role of *GAPLINC* in the biological processes of several malignancies. Its high expression enhances the migration and proliferation of renal carcinoma cell through increasing the expression of *CSF1* by sponging *miR*-135b-5p (34). It has also been suggested that overexpression of *GAPLINC* promotes invasion in colorectal cancer by targeting *SNAI2* through binding with PSF and NONO (35). Likewise, it has been disclosed that *GAPLINC* expression upregulated in human non-small cell lung cancer (NSCLC) is

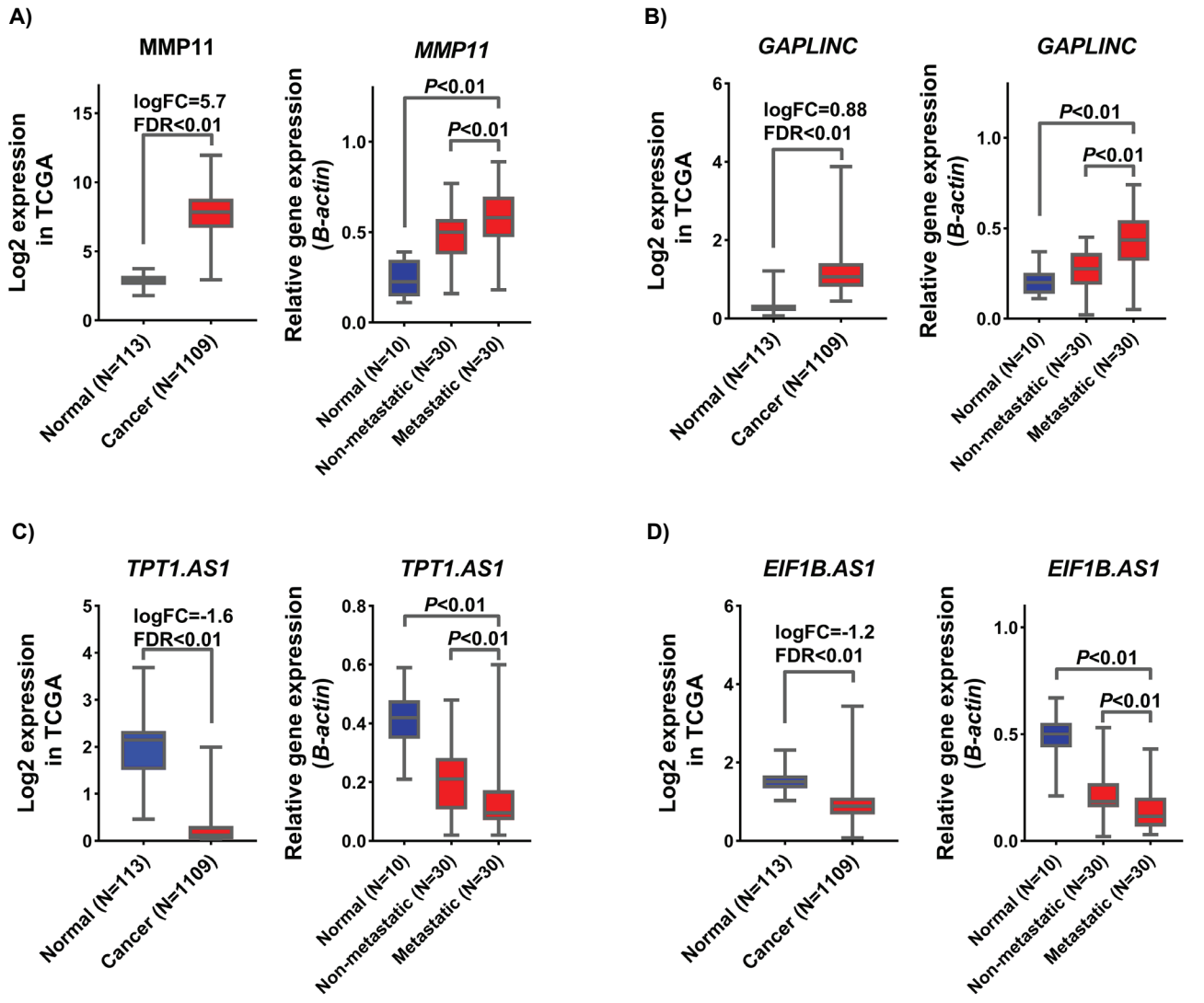

Figure 6. Comparison of expression level changes of candidate lncRNAs with the MMP11 gene. The expression level of **A**) *MMP11*, **B**) *GAPLINC*, **C**) *TPT1-AS1*, and **D**) *EIF1B-AS1* in cancer compared to normal samples using GhraphPad (V 8).

associated with poor prognosis in patients with such a disease (36). The expression pattern of *GAPLINC* reported in previous investigations is comparable to that reported in the current study. Hence, *GAPLINC* could serve as a potential prognostic and metastatic factor.

Another lncRNA candidate identified in this study is *TPT1-AS1* that has not demonstrated similar expression pattern in different types of cancers. For instance, overexpression of lncRNA *TPT1-AS1* has been shown to suppress hepatocellular carcinoma cell proliferation by downregulating *CDK2* (37), while the high expression level of *TPT1-AS1* was found to be correlated with unfavorable clinicopathological characteristics in colorectal cancer with less tumorigenesis and metastasis *in* vivo (38). Moreover, *TPT1-AS1* expression is significantly greater in liver cancer tissues and cell lines than adjacent paracancerous tissues (39).

LncRNA array and bioinformatics analysis in the present study exhibited the downregulation of *TPT1-AS1* in metastatic, but not non-metastatic, BC tissues. Therefore, further research is needed to unravel the role and mechanism of *TPT1-AS1* in tumorigenesis and

metastasis. Until now, no research has been conducted on the function of *EIF1B-AS1* in cancer disease, but our bioinformatic analysis results displayed the potential role of this lncRNA in BC. Thus, we evaluated the expression pattern of this novel lncRNA in TNBC tissues (*in vitro*); however, our findings were significant ($P<0.01$). More studies are, therefore, demanded to identify *EIF1B-AS1* as a biomarker in BC or other cancers.

The presented study successfully utilized deep learning methodologies to classify patients into non-metastatic and metastatic breast cancer cases using a new expression matrix, comprising 51 lncRNAs derived from five microarray datasets. Despite the achievement of promising results, i.e. 82% training accuracy of 82%, validation accuracy of 81%, and test accuracy of 70%, several challenges and limitations remain to be addressed. Firstly, the reliance on a specific set of lncRNAs due to platform constraints may limit the generalizability of the model to other datasets with different lncRNA profiles. In order to address the challenge posed by a vast number of features compared to the sample size, which often leads to data redundancy and overfitting, we employed a strategy leveraging our biological understanding to identify key genes and reduce the dimensionality of our dataset. We conducted standard differential expression analysis and assessed the correlation of lncRNAs used herein with well-established genes associated with BC. The specific criteria and thresholds used for feature selection are detailed in the current study. While alternative techniques such as autoencoders or machine learning methods (e.g. PCA, t-SNE, and UMAP) could be employed, we did not use these methods because autoencoders require extensive data, which could limit their effectiveness in our context. Additionally, while PCA offers dimensionality reduction, its linear nature may not adequately capture the complexity of non-linear transcriptomic data. t-SNE, primarily a visualization tool, lacks robustness in dimension reduction due to its stochastic nature. Although UMAP presents a promising alternative, it predominantly relies on data correlations and non-linear relationships, lacking integration of biological insights. In contrast, our approach by applying biological knowledge offers the advantage of informed feature selection, enhancing the interpretability and generalizability of our findings. Moreover, the performance of model could be influenced by the imbalanced distribution of non-metastatic and metastatic samples within the datasets, warranting further investigation into strategies for mitigating class imbalance. Furthermore, the interpretation of the model's decisions and the identification of key lncRNAs contributing to classification remain challenging tasks, requiring advanced techniques for model explainability and feature importance analysis. Future directions for improving deep learning methodologies in this context could involve the integration of multi-omics data sources, including genomic, transcriptomic, and epigenomic data, to enhance the robustness and predictive power of the models. Moreover, exploring novel architectures, regularization techniques, and transfer learning approaches tailored to biomedical data could facilitate the development of more accurate and interpretable models for BC metastasis prediction. Lastly, validation and external validation of the developed models on independent datasets are essential steps towards ensuring their clinical relevance and generalizability, ultimately advancing their translation into clinical practice for personalized cancer management.

6. Conclusion

The present study evaluated the role of lncRNAs, including *GAPLINC*, *TPT1-AS1*, and *EIF1B-AS1*, in the diagnosis of metastatic TNBC samples through Bioinformatics and Machine Learning. Non-coding RNAs candidates identified in this study exhibited significantly different expression levels in metastatic TNBC tissues vs. non-metastatic samples. As the candidate lncRNAs were expressed apparently, we can deduce that these could serve as novel diagnostic and prognostic biomarkers in metastatic BC. While the present findings may help better understand the crucial role of lncRNAs, identifying their mechanisms in the pathogenesis of cancers is still urgent because these RNAs have capability of activating and repressing gene expression via different mechanisms, which act alone or in combination with miRNAs and other molecules as a part of various pathways. Moreover, it is necessary to demonstrate the reliability of *GAPLINC*, *TPT1-AS1*, and *EIF1B-AS1* as indices for diagnosing other subtypes of BC or in human body fluids. Our findings represent a promising and powerful idea that we could produce pannels, measuring the expression of not all the genes, but only a few lncRNAs, and could

predict and monitor the stage and state of the patients using only these genes. With further computational resources, we would be able to reconstruct the whole expression matrix of patients and therefore make the gene regulation map of patients only using a few genes, improving our understanding of BC and its progression.

References

1. Boere I, Lok C, Poortmans P, Koppert L, Painter R, Heuvel-Eibrink MMV, Amant F. Breast cancer during pregnancy: epidemiology, phenotypes, presentation during pregnancy and therapeutic modalities. *Best Pract Res Clin Obstet Gynaecol*. 2022;**82**:46-59. doi: 10.1016/j.bpobgyn.2022.05.001
2. Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, *et al*. A population-based study of genes previously implicated in breast cancer. *N Engl J Med*. 2021;**384**(5):440-451. doi:10.1056/NEJMoa2005936
3. Colditz GA, Kaphingst KA, Hankinson SE, Rosner B. Family history and risk of breast cancer: nurses' health study. *Breast Cancer Res Treat*. 2012;**133**(3):1097-1104. doi: 10.1007/s10549-012-1985-9
4. Rathinasamy B, Velmurugan BK. Role of lncRNAs in the cancer development and progression and their regulation by various phytochemicals. *Biomed Pharmacother*. 2018;**102**:242-248. doi: 10.1016/j.biopha.2018.03.077
5. Lourenço C, Conceição F, Jerónimo C, Lamghari M, Sousa DM. Stress in metastatic breast cancer: to the bone and beyond. *Cancers* (Basel). 2022;**14**(8):1881. doi:10.3390/cancers14081881
6. González de Vega R, Clases D, Luisa Fernández-Sánchez M, Eiró N, O González L, Vizoso FJ, Doble PA, Sanz-Medel A. MMP-11 as a biomarker for metastatic breast cancer by immunohistochemical-assisted imaging mass spectrometry. *Anal Bioanal Chem*. 2019;**411**(3):639-646. doi:10.1007/s00216-018-1365-3
7. Lee K, Chng J, Jha, S. Prognostic biomarkers for breast cancer metastasis. *IntechOpen*. 2018. doi:10.5772/intechopen.80576
8. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, *et al.* Multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med.* 2004;**351**:2817-2826. doi:10.1056/NEJMoa041588
9. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, Jemal A, *et al*. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin*. 2019;**69**(5):363-385. doi:10.3322/caac.21565
10. Bridges MC, Daulagala AC, Kourtidis A. LNCcation: lncRNA localization and function. *J Cell Biol*. 2021;**220**(2):e202009045. doi:10.1083/jcb.202009045
11. Arun G, Aggarwal D, Spector DL. MALAT1 long non-coding RNA: functional implications. *Noncoding RNA*. 2020;**6**(2):22. doi:10.3390/ncrna6020022
12. Kothari C, Osseni MA, Agbo L, Ouellette G, Déraspe M, Laviolette F, Corbeil J, *et al*. Machine learning analysis identifies genes differentiating triple negative breast cancers. *Sci Rep*. 2020;**10**(1):10464 doi:10.1038/s41598-020-67525-1
13. Zhou LQ, Shen JX, Ahou JY, Hu Y, Xiao HJ. The prognostic value of m6A-related LncRNAs in patients with HNSCC: bioinformatics analysis of TCGA database. *Sci Rep*. 2022;**12**(1):579. doi:10.1038/s41598-021-04591-z
14. Thalor A, Joon HK, Singh G, Roy S, Gupta D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. *Comput Struct Biotechnol J*. 2022;**20**:1618-1631. doi:10.1016/j.csbj.2022.03.019
15. Győrffy, B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput Struct Biotechnol*. 2021;**19**:4101-4109. doi:10.1016/j.csbj.2021.07.014
16. Zaheed O, Samson J, Dean K. A bioinformatics approach to identify novel long, non-coding RNAs in breast cancer cell lines from an existing RNA-sequencing dataset. *Noncoding RNA Res*. 2020;**5**(2):48-59. doi:10.1016/j.ncrna.2020.02.004
17. Ni X, Yang H, Liu C. Identification of potential crucial genes associated with breast cancer using bioinformatics analysis and experimental verification. 2023;Available at: doi:10.21203/rs.3.rs-2457642/v1.
18. Zhang W, Shen Y, Huang H, Pan S, Jiang J, Chen W, Zhang T, *et al*. A rosetta stone for breast cancer: prognostic value and dynamic regulation of neutrophil in tumor microenvironment. *Front Immunol*.2020;**11**:1779. doi:10.3389/fimmu.2020.01779
19. Wang L, Mo C, Wang L, Cheng M. Identification of genes and pathways related to breast cancer metastasis in an integrated cohort. *Eur J Clin Invest*. 2021:**51**(7):e13525. doi:10.1111/eci.13525
20. Brancaccio M, Giachino C, Iazzetta AM, Cordone A, Marino ED, Affinito O, *et al*. Integrated bioinformatics analysis reveals novel miRNA as biomarkers associated with preeclampsia. *Genes* (Bssel). 2022;**13**(10):1781. doi:10.3390/genes13101781
21. Lee JO, Chu J, Jang G, Lee M, Chung YJ. tReasure: R-based GUI package analyzing tRNA expression profiles from small RNA sequencing data. *BMC Bioinform*. 2022;**23**(1):155. doi:10.1186/s12859-022-04691-1
22. Jafarinejad-Farsangi S, Moazzam-Jazi M, Ghale-Noie ZN, Askari N, Karam ZM, Mollazadeh S, Hadizadeh M. Investigation of genes and pathways involved in breast cancer subtypes through gene expression meta-analysis. *Gene*. 2022;**821**:146328. doi:10.1016/j.gene.2022.146328
23. Quayle LA, Spicer A, Ottewell PD, Holen I. Trans-criptomic profiling reveals novel candidate genes and signalling programs in breast cancer quiescence and dormancy. *Cancers* (Basel). 2021;**13**(16):3922. doi:10.3390/cancers13163922
24. Tavousi N, Taqizadeh Q, Nasiriyan E, Tabaeian P, Rezaei M, Azadeh M. ADAMTS5 modulates breast cancer development as a diagnostic biomarker and potential tumor suppressor, regulating by BAIAP2-AS1, VTI1B, CRNDE, and hsa-miR-135b-3p: *integrated systems biology and experimental approach*. 2022; doi:10.21203/rs.3.rs-1861409/v1.
25. Sui Y, Li S, Fu XQ, Zhao ZJ, Xing S. Bioinformatics analyses of combined databases identify shared differentially expressed genes in cancer and autoimmune disease. *J Transl Med*. 2023;**21**(1):109. doi: 10.1186/s12967-023-03943-9
26. Srivastava H, Ferrell D, Popescu GV. NetSeekR: a network analysis pipeline for RNA-Seq time series data. BMC. *Bioinformatics*. 2023;**23**(54). doi:10.1186/s12859-021-04554-1
27. Hunt GP, Grassi L, Henkin R, Smeraldi F, Spargo TP, Kabiljo R, Koks S, *et al*. GEOexplorer: a webserver for gene expression analysis and visualisation. *Nucleic Acids Res*. 2022;**50**(W1):W367-W374. doi:10.1093/nar/gkac364

28. Honda M, Iima M, Kataoka M, Fukushima Y, Ota R, Ohashi A, Toi M, Nakamoto Y. Biomarkers predictive of distant disease-free survival derived from diffusion-weighted imaging of breast cancer. *Magn Reson Med Sci*. 2022;**22**(4):469-476. doi:10.2463/mrms.mp.2022-0060
29. Lin RH, Lin CS, Chuang CL, Kujabi BK, Chen YC. Breast Cancer survival analysis model. *Appl Sci*. 2022;**12**(4):1971. doi:10.3390/app12041971
30. Sahu D, Ho SY, Juan HF, Huang HC. High-risk, expression-based prognostic long noncoding RNA signature in neuroblastoma. *JNCI Cancer Spectr*. 2018;**2**(2):pky015. doi:10.1093/jncics/pky015
31. Kotsiantis SB, Zaharakis ID, Pintelas PE. Supervised machine learning: A review of classification and combining techniques. *Artiv Intel. Rev*. 2007;**26**(3):159-190. doi:10.1007/s10462-007-9052-3
32. Mavaddat N, Pharoah PDP, Michailidiu K, Tyrer J, Brook MN, Bollaa MK. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst*. 2015;**107**(5):djv036.
33. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast. *Cancer (Dov Med Press)*. 2019;**11**:151-164. doi:10.2147/BCTT.S176070
34. Wang S, Yang X, Xie W, Fu S, Chen Q, Li Z, Zhang Z, *et al*. LncRNA GAPLINC promotes renal cell cancer tumorigenesis by targeting the miR-135b-5p/CSF1 Axis. *Front Oncol*. 2021;**11**:718532. doi:10.3389/fonc.2021.718532
35. Yang P, Chen T, Xu Z, Zhu H, Wang J, He Z. Long noncoding RNA GAPLINC promotes invasion in colorectal cancer by targeting SNAI2 through binding with PSF and NONO. *Oncotarget*. 2016;**7**(27):42183-42194. doi:10.18632/oncotarget.9741
36. Gu H, Chen J, Song Y, Shao H. Gastric adenocarcinoma predictive long intergenic non-coding RNA promotes tumor occurrence and progression in non-small cell lung cancer via regulation of the miR-661/eEF2K signaling pathway. *Cell Physiol Biochem*. 2018;**51**(5): 2136-2147. doi:10.1159/000495831
37. Wei W, Huang X, Shen X, Lian J, Chen Y, Wang W, Huang J, Zhang B. Overexpression of lncRNA TPT1-AS1 suppresses hepatocellular carcinoma cell proliferation by downregulating CDK2. *Crit Rev Eukaryot Gene Expr*.2022;**32**(1):1-9. doi:10.1615/CritRevEukaryotGeneExpr.2021039224
38. Zhang Y, Sun J, Qi Y, Wang Y, Ding Y, Wang K, Zhou Q, *et al*. Long non-coding RNA TPT1-AS1 promotes angiogenesis and metastasis of colorectal cancer through TPT1-AS1/NF90/VEGFA signaling pathway. *Aging*. 2020;**12**(7):6191-6205. doi:10.18632/aging.103016
39. Li H, Jin J, Xian J, Wang W. lncRNA TPT1AS1 knockdown inhibits liver cancer cell proliferation, migration and invasion. *Mol Med Rep*. 2021;**24**(5):782. doi:10.3892/mmr.2021.12422