参赛队员姓名：张婉玉

中学：中国人民大学附属中学

省份：北京

国家/地区：中国

指导教师姓名：倪若尧

指导教师单位：中国科学院动物研究所

论文题目：Discovery of Diagnostic Identifiers and Potential Therapeutic Targets for Triple-Negative Breast Cancer (TNBC) Using Machine Learning Methods

Discovery of Diagnostic Identifiers and Potential Therapeutic Targets for Triple-Negative Breast Cancer (TNBC) Using Machine Learning Methods

Wanyu Zhang

High School Affiliated to Renmin University of China

Abstract

Triple-Negative Breast Cancer (TNBC) has a very poor prognosis. Inaccurate diagnoses and limited treatment options call for the discovery of new identifiers and drug targets for TNBC. In this research, the expression profiles of 165 TNBC tissues and 33 normal breast tissues were obtained. Differential analysis was performed, screening out 325 differentially expressed genes (DEGs), in which 155 genes were up-regulated and 170 genes were down-regulated. The DEGs were further explored through function and pathway enrichment analyses. Three machine learning algorithms, namely Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine - Recursive Feature Elimination (SVM-RFE), and Random Forest (RF), were applied to select features that are the most representative of TNBC. The results from the three algorithms were intersected, and the predictive power of the four overlapped genes was validated. Eventually, *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18* were identified to be biomarkers and therapeutic targets for TNBC.

Key Words

Triple-Negative Breast Cancer, machine learning, Absolute Shrinkage and Selection Operator, Support Vector Machine, Random Forest, cancer detection, cancer treatment.

Abbreviations

TNBC – Triple Negative Breast Cancer
ER – Estrogen Receptor
PR – Progesterone Receptor
HER2 – Human Epidermal Growth Factor Receptor 2
DEGs – Differentially Expressed Genes
GO – Gene Ontology
KEGG – Kyoto Encyclopedia of Genes and Genomes
ML – Machine Learning
LASSO – Absolute Shrinkage and Selection Operator
SVM-RFE – Support Vector Machine - Recursive Feature Elimination
RF – Random Forest
ROC – Receiver Operating Characteristics
AUC – Area Under Curve

Table of Contents

1 Introduction

Breast cancer has become the most commonly diagnosed cancer since 2020 (Lei et al., 2021). It has a high mortality, accounting for more than 15% of cancer deaths worldwide (Lei et al., 2021). Triple-Negative Breast Cancer (TNBC) is an invasive breast cancer subtype characterized by the absence of three receptors, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Around 12%-17% of breast cancer patients have TNBC (Foulkes et al., 2010). It has a poor prognosis, with a 25% recurrence rate and a 75% mortality rate within 3 months after recurrence (Yin et al., 2020; Zhang et al., 2015).

TNBC is generally diagnosed by morphological imaging and immunohistochemistry (Penault-Llorca & Viale, 2012). Morphological imaging techniques like ultrasound examination may identify the smooth border, which is typical of some TNBC tumors. Yet the imaging techniques cannot inspect the internal fibrosis and necrosis, which are more common in TNBC tumors (Penault-Llorca & Viale, 2012). Immunohistochemistry determines the expression level of ER, PR, and HER2. It is required to further validate the diagnosis. However, immunohistochemistry suffers from inaccuracy, as small proportions of TNBC tissue may still be ER, PR, or HER2 positive (Pusztai et al., 2010). This calls for more accurate diagnostic approaches. This is where gene expression profiling analysis steps into place. In the gene expression profiling analysis, TNBC is normally classified as a subtype of basal-like breast cancer, since 60% to 90% of the genes overlap in these two types of breast cancer (Goldhirsch et al., 2013; Yin et al., 2020). According to distinct transcriptome profiles, TNBC can be further divided into six subtypes, basal-like 1 (BL1), basal-like 2 (BL2), mesenchymal (M), mesenchymal-stem-like (MSL), immunomodulatory (MI), and luminal androgen receptor (LAR) (Lehmann et al., 2011). Basal-like 1 TNBC expresses high levels of cell-cycle-related genes like *PIK3CA* and *AKT2* (Yin et al., 2020). Basal-like 2 TNBC exhibits abnormal glycolysis, gluconeogenesis, and growth factor signaling pathways and expresses elevated levels of genes *TP63* and *MME* (Lehmann et al., 2011). Mesenchymal TNBC is characterized by highly enriched cell differentiation, extracellular matrix receptor interaction, and cell motility pathways; it up-regulates genes like *ABCB1* and HOX genes (Lehmann et al., 2011; Yin et al., 2020). Mesenchymal-stem-like TNBC has an expression profile similar to mesenchymal TNBC but also up-regulates growth factor signaling pathways (Lehmann et al., 2011). Immunomodulatory TNBC, as its name indicates, enriches immune-related pathways like immune cell signaling and cytokine signaling pathways (Lehmann et al., 2011). The luminal androgen receptor subtype is the most special among the six, Its hormone regulatory pathways are highly activated. It up-regulates androgen receptor mRNAs and other genes downstream of androgen receptor

targets (Lehmann et al., 2011).

The phenotype of TNBC results in its insensitivity to endocrine therapies and most targeted therapies, which require the presence of at least one hormone receptor. Therefore, the standardized treatment for TNBC patients is chemotherapy. Commonly used chemotherapy drugs include taxanes, 5-fluorouracil, and more (Mustacchi & De Laurentiis, 2015; Yin et al., 2020). Taxanes interfere with microtubule formation, and 5-Fluorouracil's metabolites are selective against tumor cells (Mustacchi & De Laurentiis, 2015; Yin et al., 2020). These chemotherapy drugs are often used in conjugation with one another. As patients grow resistant to chemotherapy drugs, immunotherapy, targeted therapy, or radiotherapy may also be introduced. Immunotherapy mainly focuses on checkpoint inhibitors and cancer vaccines (Vikas et al., 2018), targeted therapy attacks the subtle genetic and phenotypical differences between each TNBC subtype (Yin et al., 2020), and radiotherapy facilitates local control over the tumor before and after surgery (He et al., 2018).

Unfortunately, the mentioned diagnostic methods and treatment options exhibit drawbacks. Despite the multiple phenotypical and genetic markers, misdiagnosis still happens more frequently in TNBC than in non-TNBC (Elfgen et al., 2019). Immunotherapy has high uncertainty, working differently for each patient. Targeted therapy options are limited for TNBC patients since there are few targets to attack. Radiotherapy requires a personalized design (He et al., 2018), which may burden patients and their families financially. Thus, there is an urgent need to identify novel diagnostic identifiers, which may also serve as therapeutic targets, for TNBC.

The workflow of this study is shown in **Fig 1**. In this research, transcriptome data of 165 TNBC tissues and 33 normal breast tissues were retrieved from the online database Gene Expression Omnibus (NCBI-GEO) for differential analysis. The identified up-regulated and down-regulated genes were explored through function and pathway enrichment analyses. Then, three machine-learning algorithms were used to screen for genetic identifiers from the training dataset. The models were validated using the testing dataset. The outputs of the three models were intersected to find the common genes. In the end, four critical genes, *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18*, were discovered to be identifiers and therapeutic targets for TNBC. The purpose of this research is not only to identify new biomarkers for TNBC but also to further advocate for applying the method of machine learning to cancer studies, as the development of machine learning may play a crucial role in the future discovery of biomarkers for more cancers of poor prognosis.

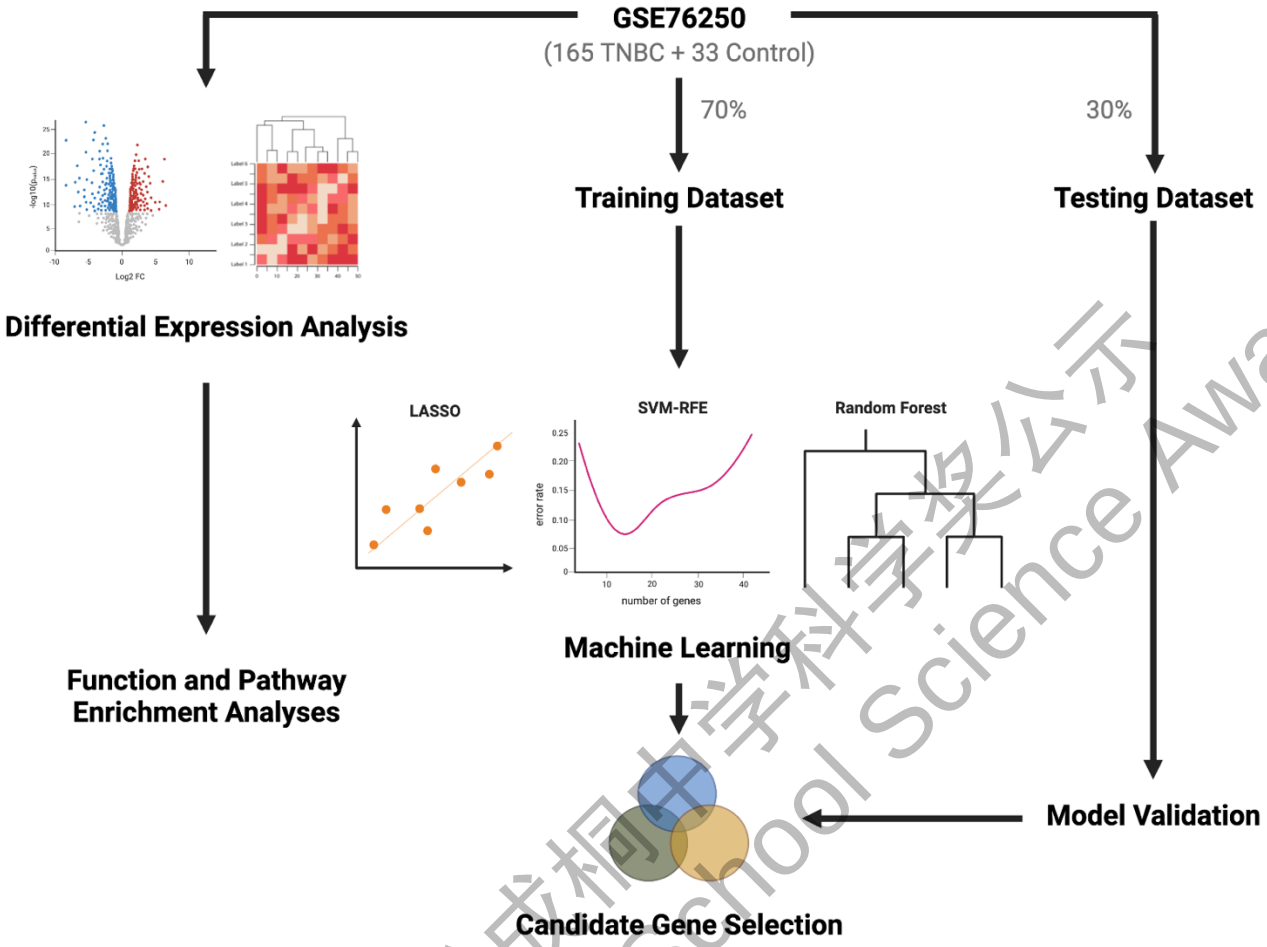

Figure 1. The general workflow of the research. The graph was created with BioRender.

2 Materials and Methods

2.1 Databases

The dataset was downloaded from Gene Expression Omnibus (NCBI-GEO), a public functional genomics data repository supported by the National Center for Biotechnology Information (NCBI) (*Home - GEO - NCBI*, n.d.).

GSE76250 was obtained by typing "TNBC homo sapiens" in the search bar. GSE76250 contains the whole expression profiles of the messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs) of 165 TNBC samples and 33 normal breast tissues. The platform was [HTA-2_0] Affymetrix Human Transcriptome Array 2.0 [transcript (gene) version].

2.2 Identifying and Filtering Differentially Expressed Genes

The expression matrix was pre-processed using the *GEOquery* package in R and went through quantile normalization by using the package *affy* (Davis & Meltzer, 2007; Gautier et al., 2004). Differential expression analysis was performed with the R package *limma*, and p-values were adjusted with the Benjamini-Hochberg method (Ritchie et al., 2015). The cut-off values were set at adjusted p-value less than 0.05 and $|\log_2(\text{Fold Change})|$ greater than 1. A gene is identified as an up-regulated gene when its adjusted p-value is less than 0.05 and $\log_2(\text{Fold Change})$ is greater than 1; a gene is recognized as a down-regulated gene when its adjusted p-value is less than 0.05 and $\log_2(\text{Fold Change})$ is less than -1.

The heatmap exhibiting the differentially expressed genes (DEGs) was generated using the R package *pheatmap*. The volcano plot was made with the R package *ggplot2*.

2.3 Function and Pathway Enrichment Analyses

Gene Ontology (GO) Function Enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Enrichment were performed for the identified DEGs by using the *clusterprofiler* package in R (Yu et al., 2012). GO Function Enrichment Analysis assigns the genes to three different groups of GO terms, Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). KEGG Pathway Enrichment Analysis sorts genes into the biological pathways they belong to. Both enrichment analyses had a cutoff value of adjusted p-values less than 0.05. The p-values were adjusted using the Benjamini-Hochberg method.

Gene Set Enrichment Analysis (GSEA) was carried out using the R package *clusterprofiler* (Yu et al., 2012). The genes were first ranked based on $\log_2$(Fold Change) in decreasing order. Then the genes were mapped into KEGG pathways, and the enrichment score for each pathway was calculated. The enrichment score increases when a gene that is in the pathway was encountered and decreases when a gene that is not in the pathway was encountered.

2.4 Machine Learning, Model Construction, Important Genes Identification, and Model Validation

The data were divided into the training group and the testing group with a ratio of 7:3. The training group contained 116 TNBC samples and 23 normal samples. The testing group contained 49 TNBC samples and 10 normal samples.

Three machine learning (ML) algorithms, Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine – Recursive Feature Elimination (SVM-RFE), and Random Forest (RF) were applied. LASSO was performed with the R package *glmnet* with 10-fold cross-validation. The response type was set as "binomial". The SVM-RFE model was constructed with the R packages *e1071* and *caret* with 10-fold cross-validation (Dimitriadou et al., 2009; Kuhn, 2008). The recursive feature elimination was performed with 50 genes and with 10-fold cross-validation. The genes were ranked according to their significance each time the model was run. The average rank of each gene is calculated. The R package *randomForest* was used to train the RF model (Liaw & Wiener, 2002). RF randomly bootstraps genes into different nodes and calculates the importance of each gene based on the ability to decrease the impurity of the node. Each gene's importance was reflected by the Gini index. The number of trees was set at 500.

Critical genes identified by LASSO, the top 40 genes ranked by SVM-RFE and RF were plotted in a Venn diagram with the R package *venn*. The intersecting genes were selected for further analysis.

The ML models were validated using the receiver operating characteristic (ROC) curve. The area under curve (AUC) and 95% confidence interval (CI) were also determined for the models. AUC indicates the usefulness of the model. An AUC higher than 0.8 means the model is very useful.

3 Results

3.1 The Identification of Differentially Expressed Genes

In the training group, 325 DEGs were identified using the cut-off value of adjusted p-value < 0.05 and $|\log_2(\text{Fold Change})| > 1$. The results were shown in **Fig. 2**. **Fig. 2A** exhibits the expression level of the DEGs in the TNBC group and the control group after quantile normalization. The colors validate that the expression levels of the DEGs were significantly different in TNBC patients and normal people, further indicating that genetic mutations may contribute to TNBC. A partial list of the DEGs is shown in **Table 1**, and the full list can be seen in **Table S1**. In **Fig. 2B**, 155 DEGs (red) with a $\log_2(\text{Fold Change})$ greater than 1 were marked as up-regulated genes; DEGs (blue) with a $\log_2(\text{Fold Change})$ less than -1 were marked as down-regulated genes. All genes with $|\log_2(\text{Fold Change})| > 2$ were labeled. The adjusted p-value was considered as well. The genes with an adjusted p-value less than 0.05 were considered to be significant DEGs. *PIP* and *MMP1* were at the margin of the volcano map (**Fig. 2B**), indicating that they were the most differentially expressed genes.

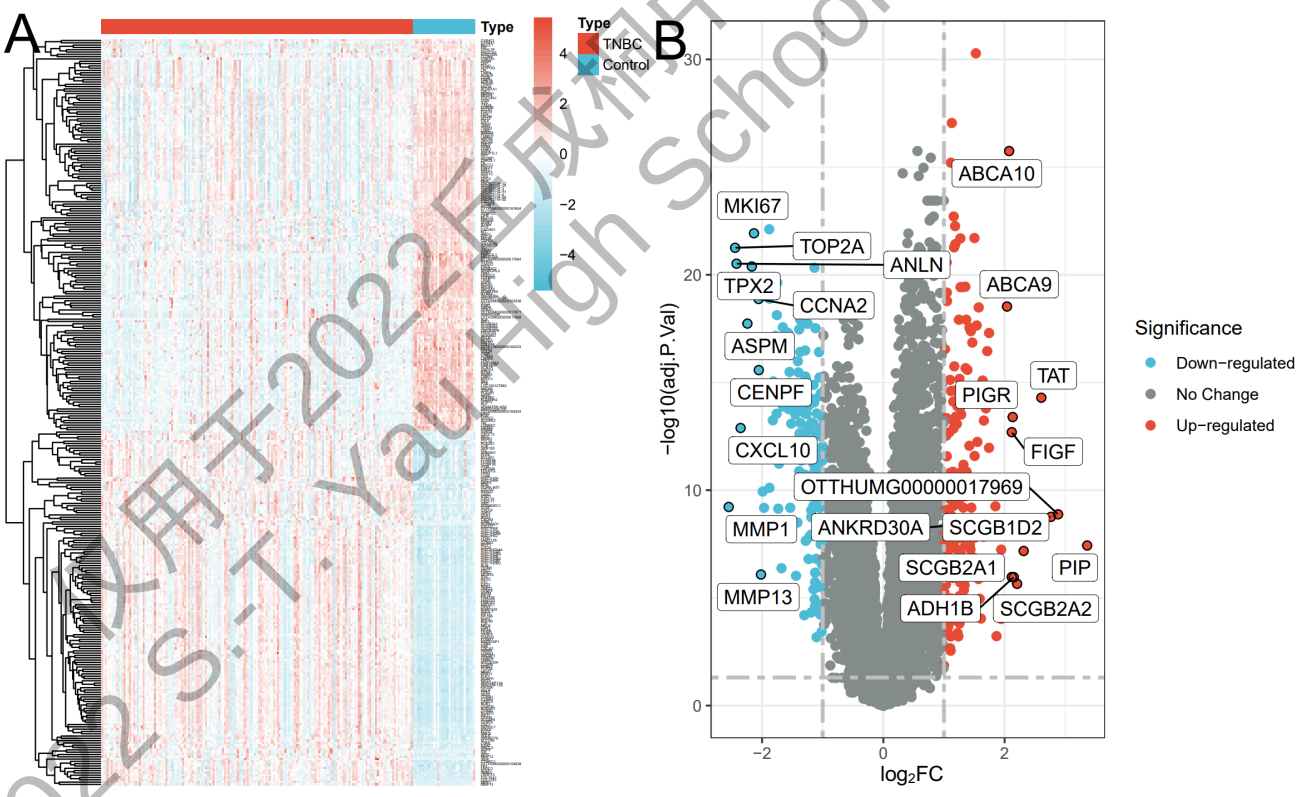

Figure 2. The differentially expressed genes in GSE76250. (A) The heatmap exhibited the expression level of the DEGs in TNBC tissues and the control group after quantile normalization. The expression level is reflected by the color bar. The more the gene was expressed in TNBC tissues compared to normal tissues, the redder the color is. The lesser the gene was expressed in TNBC tissues compared to normal tissues, the bluer the color is. **(B)** A volcano map showing the DEGs. Any DEG with a $|\log_2(\text{Fold Change})| > 2$ was labeled. Up-regulated genes are colored in red, down-regulated genes are colored in blue, and genes with no significant changes are colored in grey.

Table 1. The top 100 differentially expressed genes (50 up-regulated and 50 down-regulated) in GSE76250.

Up-regulated genes	Down-regulated genes
MMP1, TOP2A, ANLN, CXCL10, ASPM, TPX2, MKI67, CCNA2, CENPF, MMP13, ADAMDEC1, COL10A1, IF11, CXCL9, NUSAP1, DLGAP5, OTTHUMG00000154838, KIF14, NDC80, CENPE, NUF2, CEP55, HIST1H3B, SPP1, FNDC1, FAM111B, MMP12, NCAPG, TTK, PRR11, FPR3, KIF23, MELK, BUB1, CCNB2, FAM72D, CDK1, GBP5, CXCL11, PRC1, HIST1H3F, DTL, LYZ, STIL, CCL18, CKS2, OLR1, HIST1H2AB, CKAP2L, KPNA2	*MYH11, ABCA6, KIT, MUCL1, FABP4, PI15, ABCA8, PIK3C2G, ABCA9, ABCA10, SCGB2A1, FIGF, PIGR, ADH1B, SCGB2A2, SCGB1D2, TAT, ANKRD30A, OTTHUMG00000017969, PIP, CHRDL1, PROL1, FMO2, LAMA2, SCUBE2, TP63, SNORD114-1, ADH1C, SYNPO2, AK5, FGF10, SNORD114-3, ADAMTS9-AS2, TSHZ2, DMD, SNORD114-17, APOD, HPSE2, OGN, ADIPOQ, IGSF10, NTN4, OTTHUMG00000017971, CCL28, GRPR, LIFR, LRP2, SDPR, PTN, NOVA1*

3.2 Results of the Function and Pathway Enrichment Analyses

The significant DEGs then went through GO Function Enrichment Analysis, which revealed the potential function characteristics of the genes, and KEGG Pathway Enrichment Analysis, which sorted genes into different biological pathways that might be critical to disease progression.

There were 539 GO terms identified for the 325 DEGs, which are listed in **Table S2**. Most of the GO terms enriched in up-regulated genes were associated with the extracellular matrix and systematic development, including collagen-containing extracellular matrix, urogenital system development, and extracellular matrix structural constituent (**Fig. 3A**). The top 3 enriched GO terms for the down-regulated genes were closely related to cell replication, including nuclear division, organelle fission, and chromosome segregation (**Fig. 3B**).

Twenty KEGG pathways were enriched. The full list is provided in **Table S2**, while the top 5 enriched KEGG pathways are exhibited in **Fig. 3C** (for up-regulated DEGs) and **Fig. 3D** (for down-regulated DEGs). The most enriched pathways for up-regulated genes were phosphoinositide 3-kinase – protein kinase B (PI3K-AKT) signaling pathway followed by extracellular-matrix-receptor interaction and peroxisome proliferator-activated receptor (PPAR) signaling pathway (**Fig. 3C**). The top 3 enriched pathways for down-regulated genes were cell cycle, cytokine-cytokine receptor interaction, and viral protein interaction with cytokine and cytokine receptor (**Fig. 3D**).

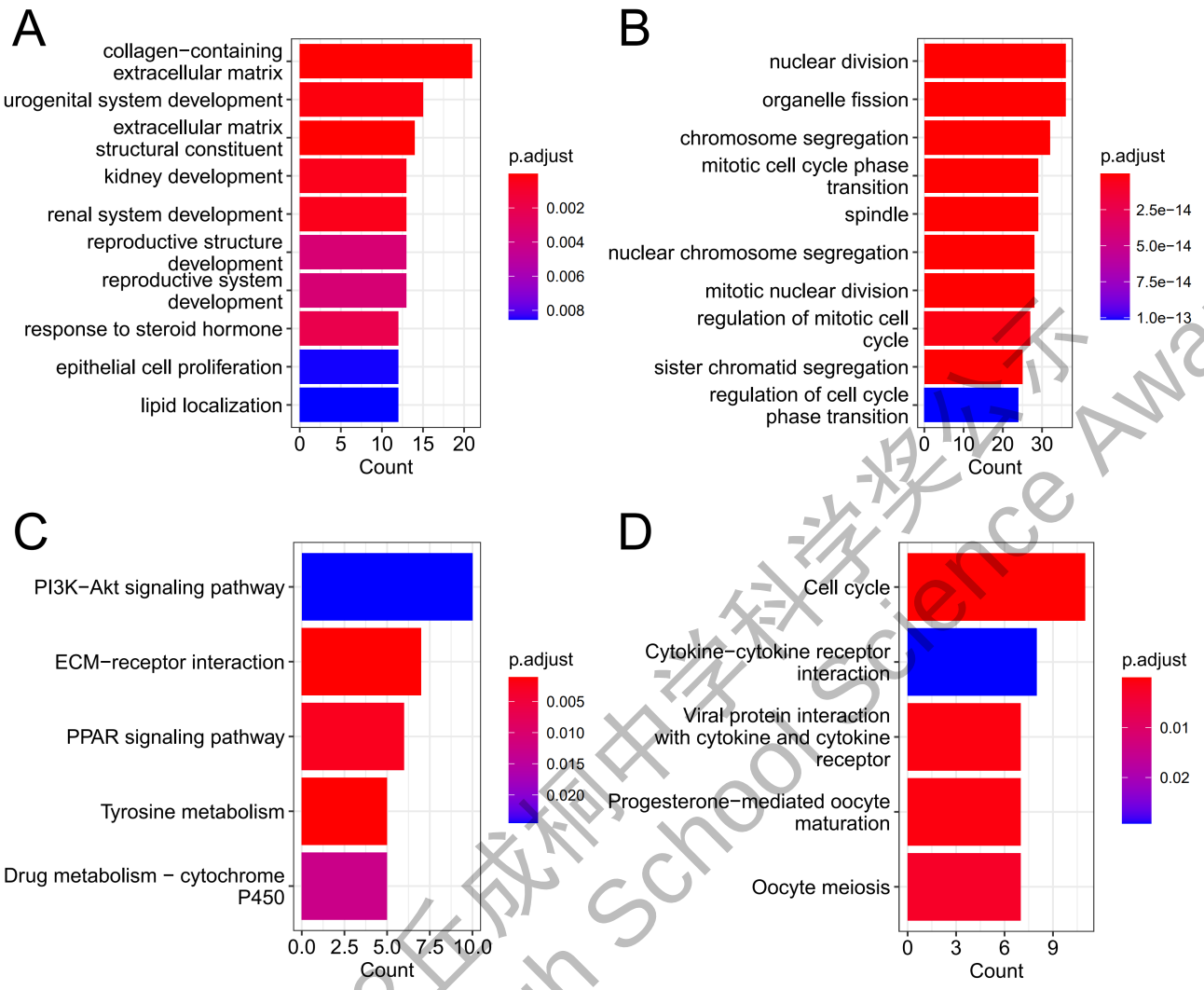

Figure 3. The results of GO and KEGG enrichment analysis. (A) Top 10 GO terms enriched for up-regulated genes. **(B)** Top 10 GO terms enriched for down-regulated genes. **(C)** Top 5 KEGG pathways enriched for up-regulated genes. **(D)** Top 5 KEGG pathways enriched for down-regulated genes.

GSEA was then performed for the full expression data. The results are shown in **Fig. 4**. The enrichment score increases when the gene encountered is in the pathway and decreases when the gene encountered is not in the pathway. The up-regulated pathways with top 5 enrichment scores are exhibited in **Fig. 4A**; the exhibited pathways were the calcium signaling pathway, cAMP signaling pathway, chemical carcinogenesis – receptor activation, mitogen-activated protein kinase (MAPK) signaling pathway, and olfactory transduction. The down-regulated pathways with top 5 enrichment scores, shown in **Fig. 4B**, included amyotrophic lateral sclerosis, human T-cell leukemia virus 1 infection, Huntington's disease, Parkinson's disease, and prion disease. The pathways with the highest and the lowest enrichment scores were the calcium signaling pathway and prion disease, respectively, indicating these two pathways function critically in TNBC pathology.

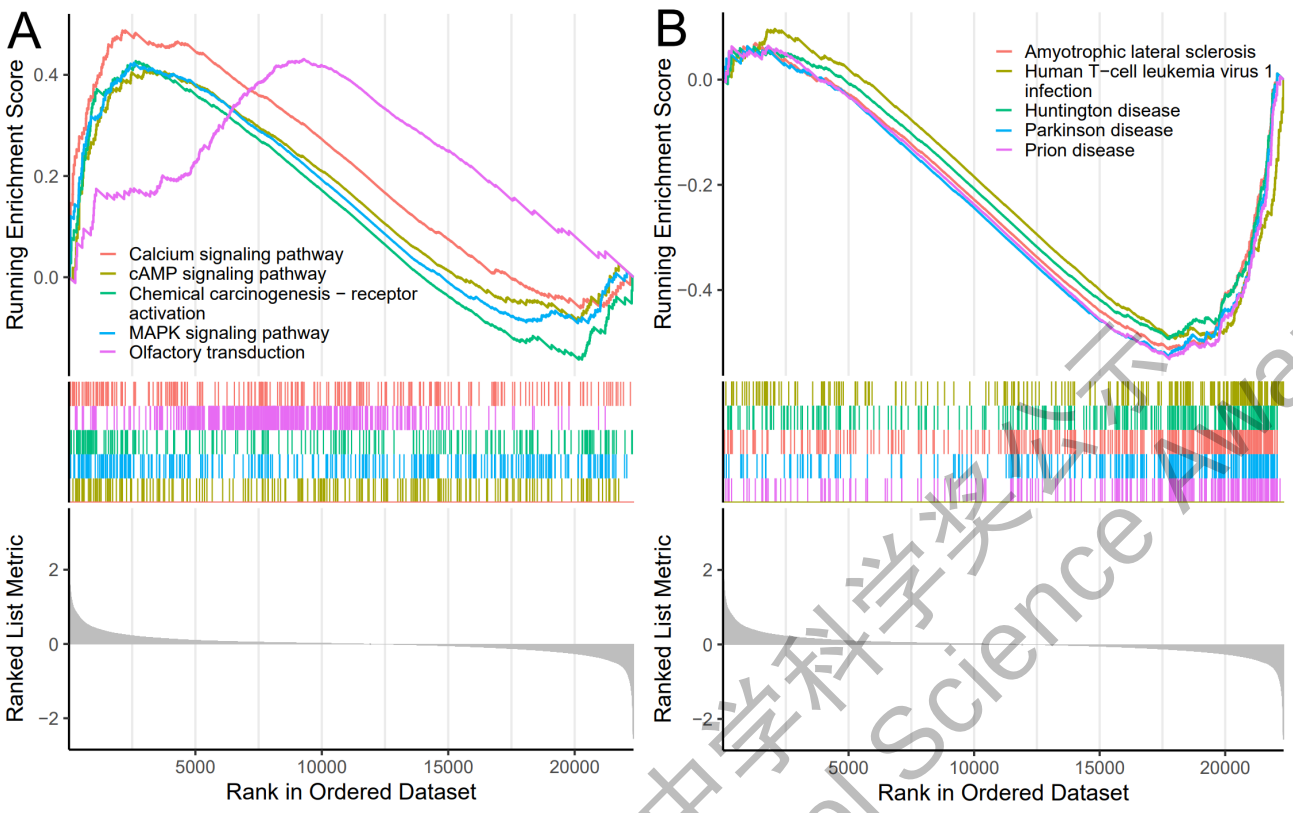

Figure 4. The results of GSEA enrichment analysis. (A) Top 5 enriched pathways for up-regulated genes. **(B)** Top 5 enriched pathways for down-regulated genes. In both **(A)** and **(B)**, genes were ranked from left to right based on their $\log_2$(Fold Change) ($\log_2$(Fold Change) reflects the level of differential expression). The gene with the highest $\log_2$(Fold Change) was on the far left. The enrichment score increases as a gene that is in the pathway was encountered, and the enrichment score decreases as a gene that is not in the pathway was encountered.

3.3 Machine Learning and Potential Biomarker Selection

Three models were successfully constructed using three machine-learning algorithms, LASSO, SVM-RFE, and RF (**Fig. 5**).

In the model created using LASSO, lambda (λ) is the regularization parameter that controls the penalty. **Fig. 5A** exhibits the binominal deviance in response to each of the lambda values tested. The binomial deviance indicates the misclassification error. The lower the binomial deviance is, the more predictive power the model should have. The lambda for the model with the lowest misclassification rate was 0.0130826. In Fig 5A, the vertical dotted line on the left shows the lambda value with the lowest error rate; the vertical dotted line on the right is the highest lambda value of the model that is within one standard error from the lambda of the optimal model. Eventually, 12 genes with coefficients not equal to zero after penalization were selected to be candidate identifiers (**Table 2**).

The SVM-RFE model was dedicated to finding a statistical separation between the expression level of certain genes in TNBC tissues and normal tissues. The model returned the average rank of the gene in the 10-fold cross-validation (**Table S3**). The higher the average rank is, the more predictive power the gene should have. The model also returned the error rate when 1 to 50 genes were incorporated. It was calculated that the error rate was the lowest when 40 genes were used in distinguishing TNBC tissues from normal tissues (**Fig. 5B**). Therefore, genes with the top 40 average rank were selected from the SVM-RFE model.

The third model was constructed using the algorithm RF. The model returned the error rate when the data was assigned to different numbers of trees. In **Fig. 5C**, the black curve reflected the overall out-of-bag (OOB) error, and the green and red curves reflected the out-of-bag error when using the trees to classify the control group and the TNBC group, respectively. A low out-of-bag error is indicative of a better performance. The overall out-of-bag error was the lowest when there were 27 trees (**Fig. 5C**). The importance of each gene was reflected by the Gini index. The Gini index reflected impurity. The more the Gini index of a gene decreased at each split, or the more impurity a gene is able to decrease, the more important the gene should be. The importance of the genes is shown in **Fig. 5D**. Genes with the top 40 importance were selected.

The 12 genes from the LASSO model, the 40 genes from the SVM-RFE model, and the 40 genes from the RF model were intersected (**Fig. 6**). There were 4 genes shared by all three models, 4 genes shared by the LASSO and the SVM-RFE models, 2 genes shared by the SVM-RFE and RF models, and 1 gene shared by the LASSO and RF models. The full list of the candidate biomarker genes is provided in **Table 2**, and the common genes are in bold font.

Table 2. The genes selected by the three machine learning algorithms. The intersecting genes are displayed in bold font.

LASSO	SVF-RFE	RF
CCL18, ***HIST1H3H****,* ***CKAP2****,* ***IL18****, MUC15, PKHD1L1,* ***ESR1****, OTTHUMG00000164324, HMGCS2, OTTHUMG00000017664, FAM196B, PROL1*	*HMGCS2, LYZ,* ***ESR1****,* ***IL18****,* ***HIST1H3H****, PKHD1L1, NCAPH, TOP2A, CCL18, MKI67, MMP12, SHCBP1, TPX2, BGN, HIST1H3B, ABCA6, IL33, FCER1G, MELK,* ***CKAP2****, MMP13, CD86, SDPR, CYP4Z1, ADH1B, TICRR, RGS1, CNN1, KIF18A, ADAMDEC1, NEK2,*	*AS2,* ***CKAP2****, AASS, ARHGAP11B, ARHGAP11A, CDKN3, NDC80, HLF, FAM189A2, CENPE, CKAP2L, HIST1H2AB, ASPM,* ***HIST1H3H****, FOXM1, CEP55, PLA2G7, NTN4, FCGR1B, SPRY2, MMP1, ANLN, DLGAP5, HIST2H3A, CACHD1, MCM10, PLEKHH2, KIF14, HIST1H3J, TGFBR3,* ***ESR1****,*

	OTTHUMG00000154838, MUC15, ANLN, BUB1, PLK1, INHBA, ABCA8, LRP2, CHL1	*OTTHUMG00000017664, MIR4524A, HIST1H2AG, CASC5, **IL18**, PLK1, HMMR, KIF2C, TP63*

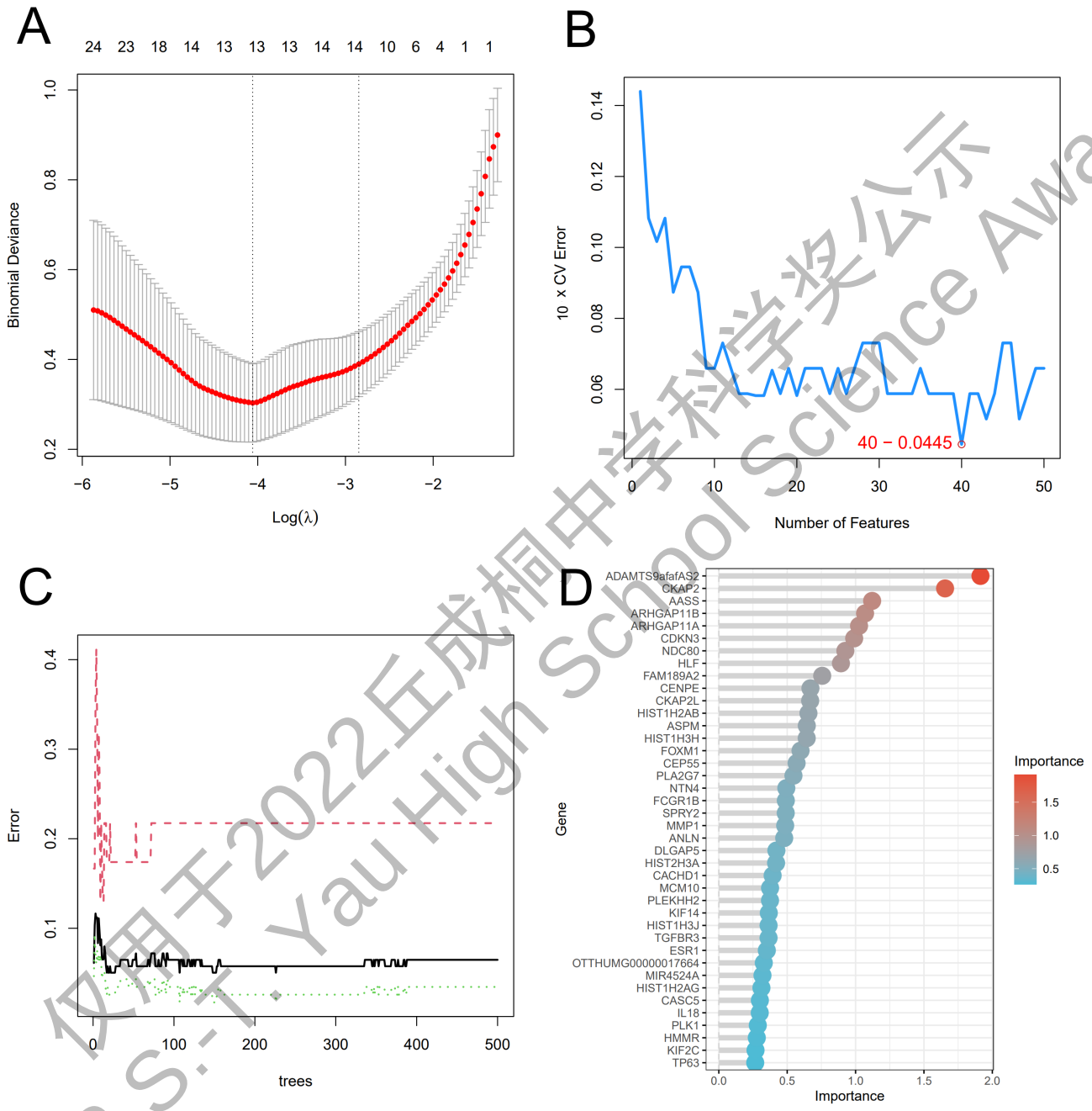

Figure 5. The results of the three machine learning algorithms. (A) The model constructed with LASSO. The vertical dotted line on the left is the lambda (λ) value with the lowest binomial deviance. The vertical dotted line on the right is the λ value that is within one standard error from the best λ value. Twelve genes were eventually selected. **(B)** The model constructed with SVM-RFE. The top 40 genes were selected, as the error rate was the lowest when 40 genes were incorporated into the model. **(C)** The model constructed with RF. The black, green, and red curves indicate the out-of-bag (OOB) error for the overall performance, the control group, and the TNBC group, respectively. The out-of-bag error rate was the lowest when there were 27 trees. Genes with the top 40 importance were selected. **(D)** The top 40 important genes from the RF model, ranked by importance calculated from the Gini index.

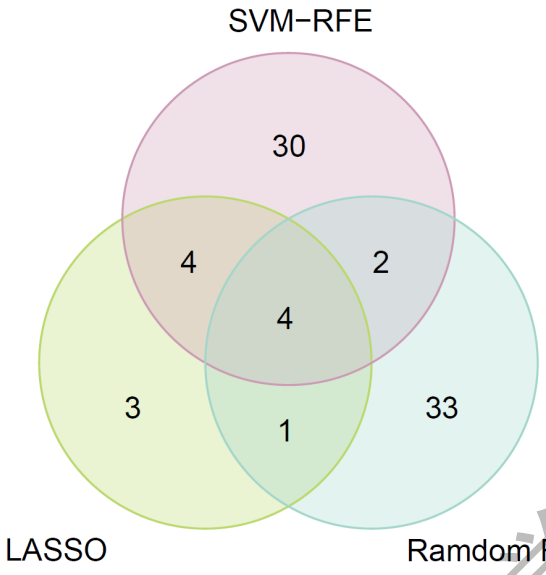

Figure 6. The common genes in the output of the three models, LASSO, SVM-RFE, and RF.

3.4 Model Validation

The 4 common genes for the LASSO, SVM-RFE, and RF models were selected for further validation.

The expression level of the four critical genes was compared in TNBC tissues and normal tissues (**Fig. 7**). The boxplots show that all four genes were very differently expressed. *CKAP2*, *HIST1H3H*, and *IL18* were up-regulated (**Fig. 7A, 7B, 7D**), and *ESR1* was down-regulated (**Fig. 7C**).

The models were validated with the receiver operating characteristics (ROC) curves. The curves were generated for both the training dataset and the testing dataset. The more a ROC curve approaches the upper left corner, the better the performance of the model. As shown in **Fig. 8**, in both the training dataset and testing dataset, the ROC curves for all four genes were near the upper-right corner, indicating that the models have high sensitivity (true positive rates) and low specificity (false positive rates). The area under curve (AUC) was calculated for each model as well. A model is considered good when its AUC exceeds 0.8. The AUC value was greater than 0.8 for all the models constructed using either the training dataset or the testing dataset, which proved that the models were predictive. Moreover, although the training group always exhibited higher AUC than the testing group for the same gene, the differences were not large. The testing dataset performed almost as well as the training group, suggesting that the four genes have good and consistent predictive powers.

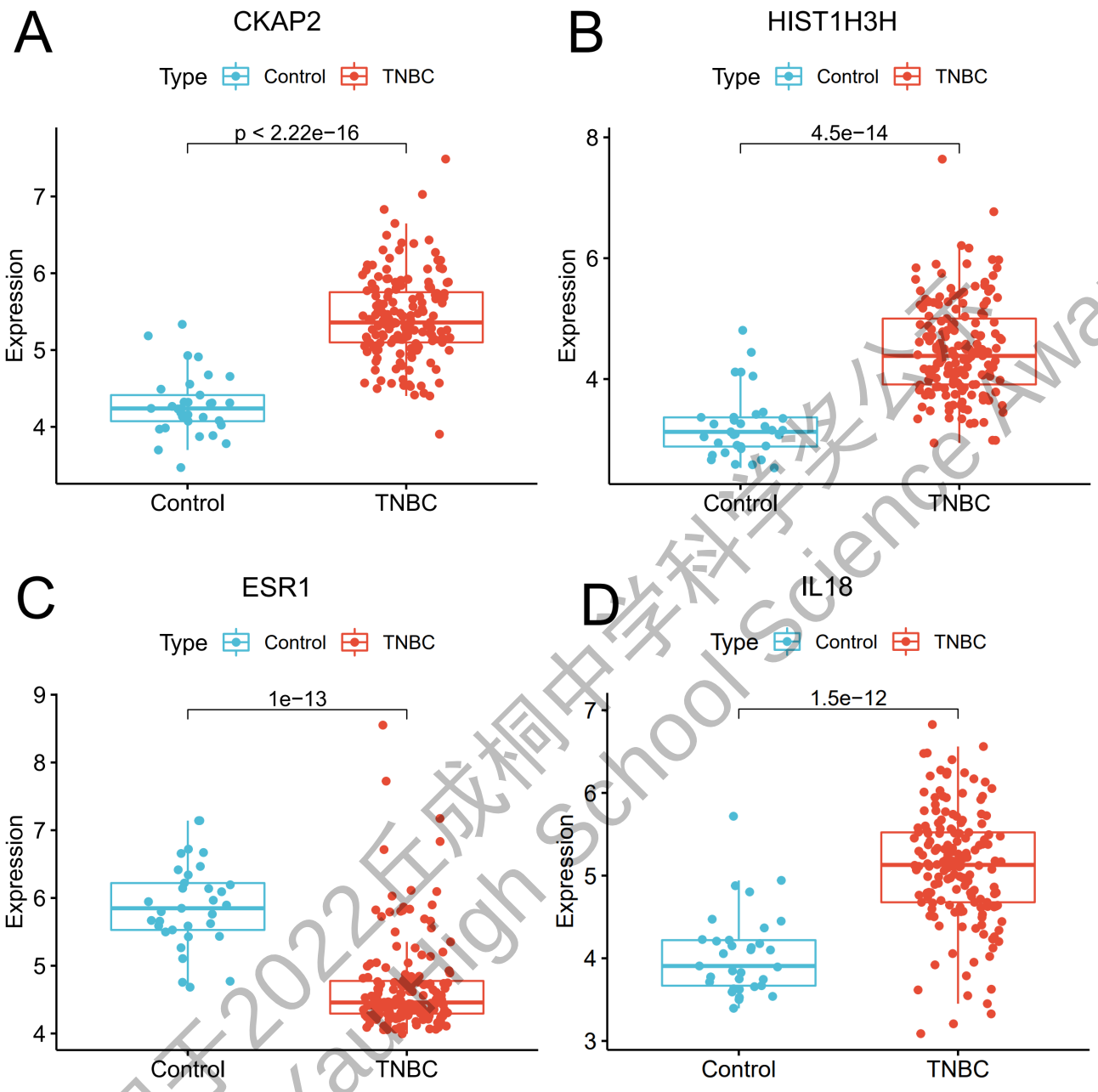

Figure 7. The expression level of the 4 critical genes in TNBC tissues and normal breast tissues (the controls). (A) *CKAP2*. **(B)** *HIST1H3H*. **(C)** *ESR1*. **(D)** *IL18*.

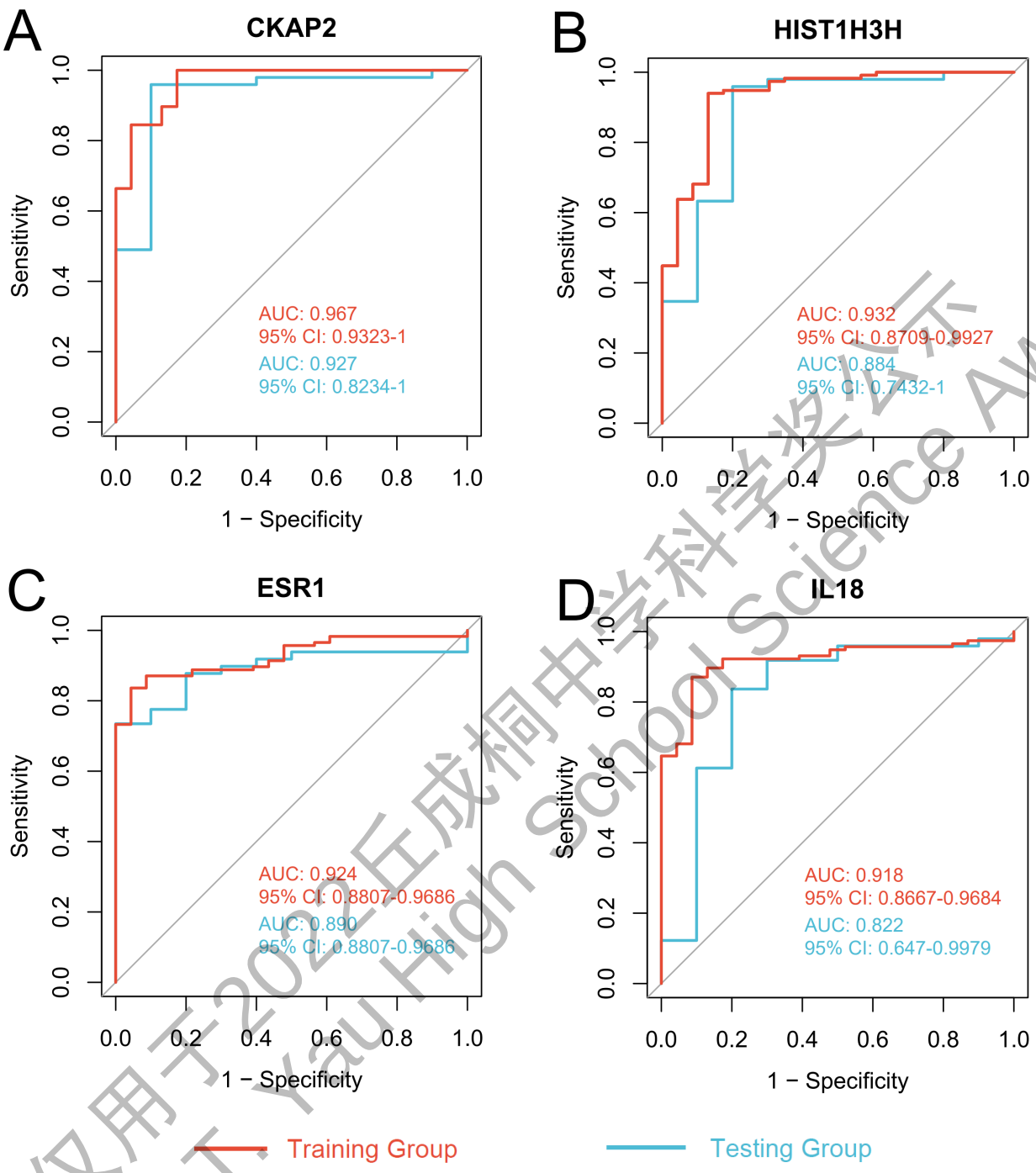

Figure 8. The ROC curve for the four critical genes. (A) *CKAP2*. **(B)** *HIST1H3H*. **(C)** *ESR1*. **(D)** *IL18*. For each gene, the curves of the training group and the testing group were mapped on the same axis. The y-axis represents sensitivity, the true positive rate. The x-axis represents specificity, the false positive rate. The 95% confidential intervals (95% CI) were calculated and presented in the graphs as well.

4 Discussion

In this research, the expression profile of 165 TNBC tissues and 33 normal breast tissues were extracted from the dataset GSE76250. Differential gene expression analysis revealed 325 DEGs. Three machine-learning algorithms, namely LASSO, SVM-RFE, and RF, were applied. LASSO outputted 12 genes, SVM-RFE outputted 40 genes, and RF outputted 40 genes. Four genes, *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18*, were common outputs of the three machine learning algorithms mentioned above (**Table 2**; **Fig. 6**). The predictive power of the four genes was then validated by ROC and AUC. The average AUC of the training group and the testing group for *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18* was 0.947, 0.908, 0.907, and 0.870, respectively.

The enrichment analyses identified several pathways that need further exploration. In the top 3 functions enriched for the up-regulated DEGs, 2 of them were related to extracellular matrix (ECM). ECM is actively involved in the progression of breast cancer. The matrix metalloproteinases degrade ECM proteins to facilitate metastasis, the integrins and other enzymes on the surface of ECM enable cancer development, and stromal cells aid in constructing blood vessels for tumors (Jena & Janjanam, 2018). This could explain the up-regulation of ECM-related functions in TNBC. Surprisingly, the top 3 enriched functions for down-regulated genes were nuclear division, organelle fission, and chromosome segregation. These functions are expected to be up-regulated instead of down-regulated, as cancer cells divide rapidly and constantly. Several pieces of research also identified nuclear division and organelle fission as top enriched functions; yet in these researches, the functions were enriched for up-regulated genes (Chen et al., 2022; Suo et al., 2020). Individual genes in the down-regulated functions were examined. Unfortunately, aside from *BRIP1*, a tumor suppressor gene that repairs DNA damage, *HORMAD1*, a gene encoding for a meiosis-specific protein, and *KIF14*, which delays the transition from metaphase to anaphase, all the other down-regulated genes were shown by previous studies to be up-regulated (Hung et al., 2013; Khan & Khan, 2021; Liu et al., 2020). Therefore, further investigations are needed to explain the abnormal down-regulation of these functions. As shown by the results of the KEGG pathway enrichment analysis, the top regulated pathway was the PI3K-AKT signaling pathway, whose abnormality is very common in all subtypes of breast cancer (LoRusso, 2016). The overexpression of upstream regulators of this pathway can lead to the progression of TNBC (Costa et al., 2018). Other enriched up-regulated pathways include the PPAR signaling pathway, which eventually activates the transcription factor PPAR-α and hence facilitates cancer development (Kwong et al., 2019). Cell cycle was the most down-regulated KEGG pathway, coinciding with the function enrichment analysis results.

The four significant genes were further examined. *CKAP2* encodes for cytoskeleton-associated protein 2. The overexpression of *CKAP2* should stabilize microtubules, leading to disrupted mitosis, abnormal cytokinesis, cell cycle arrest, and apoptosis (Tsuchihara et al., 2005). *CKAP2* has already been identified as an indicator of breast cancer. A study by Kim et al. showed that *CKAP2* can serve as an independent prognostic indicator of relapse-free survival in breast cancer patients, as *CKAP2*-positive cell count was strongly correlated with poor survival (Kim et al., 2014). Moreover, dos Santos et al. showed that the proliferation, migration, and aggregation of breast cancer cells were attenuated by knocking off *CKAP2* (dos Santos et al., 2022), which proved that *CKAP2* might be a good target in treating breast cancer. *HIST1H3H* belongs to a small cluster of histone genes, which encodes for a type of histone in the H3 family. It was also reported to be one of the prognostic predictors of breast cancer (Xie et al., 2019). *ESR1* encodes for estrogen receptor 1. The down-regulation of *ESR1* observed in this study is reasonable, as TNBC tissue does not express estrogen receptors. *ESR1* was amplified in over 20% of breast cancers (Holst et al., 2007). *IL18* encodes for interleukin-18 (IL-18), a type of pre-inflammatory cytokine. IL-18 was reported to be immunosuppressive, facilitating metastasis by up-regulating the expression of the immune checkpoint PD-1 (Terme et al., 2011). In breast cancer cells, *IL18* expression was also up-regulated by the hormone leptin via PI3K-AKT/ATF-2 signaling pathways, which eventually lead to metastasis (Li et al., 2016). Although *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18* were all proven to play a role in breast cancer identification or development, limited research was done specifically for TNBC. Kim et al. noted in their research that the predictive power of *CKAP2*-positive cell count was significant for TNBC (Kim et al., 2014), and the results of this research coincided with this finding. Aside from this, none of the other three genes was reported to have a special function in TNBC. This paper not only validated, from a machine-learning perspective, that *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18* are important for breast cancer but also proposed the four genes to be identifiers and therapeutic targets for TNBC specifically.

People are constantly looking forward to finding approaches to diagnose cancer early and to determine the prognosis of cancer accurately. ML is certainly one of the most popular ways to achieve the goals. A great amount of data have been collected over the years in virtue of advances in technologies, and ML, capable of identifying patterns in complex datasets, has become the ideal data-processing tool (Kourou et al., 2015). Additionally, the error rate of ML decreases as more and more correctly pre-processed datasets were fed in, making classification and feature selection more accurate. This characteristic made ML more useful in future cancer studies, as the datasets available for training will increase significantly in the foreseeable future.

Despite the encouraging discoveries, several limitations were present in this study. Firstly, the dataset was imbalanced; the number of TNBC tissues was five times more than the normal breast tissues that served as the control group. The sample size of the control group should be expanded using algorithms like Synthetic Minority Oversampling Technique (SMOTE) to improve the performances of the machine learning algorithms. Secondly, the three ML algorithms each exhibit their own limitations. When a group of genes is colinear, LASSO tends to select only one gene from the group and eliminate all others. SVM-RFE is very sensitive to noises, and small errors may lead to very poor performance. RF performs badly when the data is imbalanced, which unfortunately happened in this study. Although the error rate was to an extent reduced by accepting only intersecting genes, the results should be viewed critically as errors may still exist due to the limitations of the three algorithms. Thirdly, all experiments were conducted *in silico*. *In vitro* and even *in vivo* research is needed to further validate the findings. Lastly, the study failed to consider other factors that may affect the genotype; the not-considered features included but were not limited to patients' age, biological sex, ethnicity, and medical history. Future studies are expected to map more thorough and representative profiles for patients of all kinds.

5 Conclusion

In this study, through performing differential gene expression analysis and conducting function and pathway enrichment analyses, DEGs of TNBC were found to relate to multiple functions and pathways regarding the extracellular matrix and cell cycle. By further applying three machine learning algorithms, four crucial genes, namely *CKAP2*, *HIST1H3H*, *ESR1*, and *IL18*, were identified to be potential diagnostic identifiers and therapeutic targets of TNBC.

Reference

Chen, D.-L., Cai, J.-H., & Wang, C. C. N. (2022). Identification of Key Prognostic Genes of Triple Negative Breast Cancer by LASSO-Based Machine Learning and Bioinformatics Analysis. *Genes*, *13*(5), 902. https://doi.org/10.3390/genes13050902

Costa, R. L. B., Han, H. S., & Gradishar, W. J. (2018). Targeting the PI3K/AKT/mTOR pathway in triple-negative breast cancer: A review. *Breast Cancer Research and Treatment*, *169*(3), 397–406. https://doi.org/10.1007/s10549-018-4697-y

Davis, S., & Meltzer, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2009). E1071: Misc Functions of the Department of Statistics (E1071), TU Wien. In *R package version 1.5-24* (Vol. 1).

dos Santos, A., Ouellette, G., Diorio, C., Elowe, S., & Durocher, F. (2022). Knockdown of CKAP2 Inhibits Proliferation, Migration, and Aggregate Formation in Aggressive Breast Cancer. *Cancers*, *14*(15), 3759. https://doi.org/10.3390/cancers14153759

Elfgen, C., Varga, Z., Reeve, K., Moskovszky, L., Bjelic-Radisic, V., Tausch, C., & Güth, U. (2019). The impact of distinct triple-negative breast cancer subtypes on misdiagnosis and diagnostic delay. *Breast Cancer Research and Treatment*, *177*(1), 67–75. https://doi.org/10.1007/s10549-019-05298-6

Foulkes, W. D., Smith, I. E., & Reis-Filho, J. S. (2010). Triple-Negative Breast Cancer. *New England Journal of Medicine*, *363*(20), 1938–1948. https://doi.org/10.1056/NEJMra1001389

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405

Goldhirsch, A., Winer, E. P., Coates, A. S., Gelber, R. D., Piccart-Gebhart, M., Thürlimann, B., Senn, H.-J., & Panel members. (2013). Personalizing the treatment of women with early breast cancer: Highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *24*(9), 2206–2223. https://doi.org/10.1093/annonc/mdt303

He, M. Y., Rancoule, C., Rehailia-Blanchard, A., Espenel, S., Trone, J.-C., Bernichon, E., Guillaume, E., Vallard, A., & Magné, N. (2018). Radiotherapy in triple-negative breast cancer: Current situation and upcoming strategies. *Critical Reviews in Oncology/Hematology*, *131*, 96–101.

https://doi.org/10.1016/j.critrevonc.2018.09.004

Holst, F., Stahl, P. R., Ruiz, C., Hellwinkel, O., Jehan, Z., Wendland, M., Lebeau, A., Terracciano, L., Al-Kuraya, K., Jänicke, F., Sauter, G., & Simon, R. (2007). Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nature Genetics*, *39*(5), 655–660. https://doi.org/10.1038/ng2006

Home—GEO - NCBI. (n.d.). Retrieved May 19, 2022, from https://www.ncbi.nlm.nih.gov/geo/

Hung, P.-F., Hong, T.-M., Hsu, Y.-C., Chen, H.-Y., Chang, Y.-L., Wu, C.-T., Chang, G.-C., Jou, Y.-S., Pan, S.-H., & Yang, P.-C. (2013). The Motor Protein KIF14 Inhibits Tumor Growth and Cancer Metastasis in Lung Adenocarcinoma. *PLoS ONE*, *8*(4), e61664. https://doi.org/10.1371/journal.pone.0061664

Jena, M. K., & Janjanam, J. (2018). Role of extracellular matrix in breast cancer development: A brief update. *F1000Research*, *7*, 274. https://doi.org/10.12688/f1000research.14133.2

Khan, U., & Khan, Md. S. (2021). Prognostic Value Estimation of BRIP1 in Breast Cancer by Exploiting Transcriptomics Data Through Bioinformatics Approaches. *Bioinformatics and Biology Insights*, *15*, 11779322211055892. https://doi.org/10.1177/11779322211055892

Kim, H.-S., Koh, J.-S., Choi, Y.-B., Ro, J., Kim, H.-K., Kim, M.-K., Nam, B.-H., Kim, K.-T., Chandra, V., Seol, H.-S., Noh, W.-C., Kim, E.-K., Park, J., Bae, C.-D., & Hong, K.-M. (2014). Chromatin CKAP2, a New Proliferation Marker, as Independent Prognostic Indicator in Breast Cancer. *PLOS ONE*, *9*(6), e98160. https://doi.org/10.1371/journal.pone.0098160

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Kwong, S. C., Jamil, A. H. A., Rhodes, A., Taib, N. A., & Chung, I. (2019). Metabolic role of fatty acid binding protein 7 in mediating triple-negative breast cancer cell death via PPAR-α signaling. *Journal of Lipid Research*, *60*(11), 1807–1817. https://doi.org/10.1194/jlr.M092379

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, *121*(7), 2750–2767. https://doi.org/10.1172/JCI45014

Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., Zeng, H., Zhou, J., & Wei, W. (2021). Global patterns

of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Communications*, *41*(11), 1183–1194. https://doi.org/10.1002/cac2.12207

Li, K., Wei, L., Huang, Y., Wu, Y., Su, M., Pang, X., Wang, N., Ji, F., Zhong, C., & Chen, T. (2016). Leptin promotes breast cancer cell migration and invasion via IL-18 expression and secretion. *International Journal of Oncology*, *48*(6), 2479–2487. https://doi.org/10.3892/ijo.2016.3483

Liu, K., Wang, Y., Zhu, Q., Li, P., Chen, J., Tang, Z., Shen, Y., Cheng, X., Lu, L.-Y., & Liu, Y. (2020). Aberrantly expressed HORMAD1 disrupts nuclear localization of MCM8–MCM9 complex and compromises DNA mismatch repair in cancer cells. *Cell Death & Disease*, *11*(7), 1–15. https://doi.org/10.1038/s41419-020-2736-1

LoRusso, P. M. (2016). Inhibition of the PI3K/AKT/mTOR Pathway in Solid Tumors. *Journal of Clinical Oncology*, *34*(31), 3803–3815. https://doi.org/10.1200/JCO.2014.59.0018

Mustacchi, G., & De Laurentiis, M. (2015). The role of taxanes in triple-negative breast cancer: Literature review. *Drug Design, Development and Therapy*, *9*, 4303–4318. https://doi.org/10.2147/DDDT.S86105

Penault-Llorca, F., & Viale, G. (2012). Pathological and molecular diagnosis of triple-negative breast cancer: A clinical perspective. *Annals of Oncology*, *23*, vi19–vi22. https://doi.org/10.1093/annonc/mds190

Pusztai, L., Viale, G., Kelly, C. M., & Hudis, C. A. (2010). Estrogen and HER-2 Receptor Discordance Between Primary Breast Cancer and Metastasis. *The Oncologist*, *15*(11), 1164–1168. https://doi.org/10.1634/theoncologist.2010-0059

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Robinson, D. R., Wu, Y.-M., Vats, P., Su, F., Lonigro, R. J., Cao, X., Kalyana-Sundaram, S., Wang, R., Ning, Y., Hodges, L., Gursky, A., Siddiqui, J., Tomlins, S. A., Roychowdhury, S., Pienta, K. J., Kim, S. Y., Roberts, J. S., Rae, J. M., Van Poznak, C. H., … Chinnaiyan, A. M. (2013). Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature Genetics*, *45*(12), 1446–1451. https://doi.org/10.1038/ng.2823

Suo, H., Tao, Z., Zhang, L., Jin, Z., Li, X., Ma, W., Wang, Z., Qiu, Y., Jin, F., Chen, B., & Cao, Y. (2020). Coexpression Network Analysis of Genes Related to the Characteristics of Tumor Stemness in Triple-Negative Breast Cancer. *BioMed Research International*, *2020*, e7575862. https://doi.org/10.1155/2020/7575862

Terme, M., Ullrich, E., Aymeric, L., Meinhardt, K., Desbois, M., Delahaye, N., Viaud, S., Ryffel, B., Yagita, H., Kaplanski, G., Prévost-Blondel, A., Kato, M., Schultze, J. L., Tartour, E., Kroemer, G., Chaput, N., & Zitvogel, L. (2011). IL-18 Induces PD-1–Dependent Immunosuppression in Cancer. *Cancer Research*, *71*(16), 5393–5399. https://doi.org/10.1158/0008-5472.CAN-11-0993

Tsuchihara, K., Lapin, V., Bakal, C., Okada, H., Brown, L., Hirota-Tsuchihara, M., Zaugg, K., Ho, A., Itie-YouTen, A., Harris-Brandts, M., Rottapel, R., Richardson, C. D., Benchimol, S., & Mak, T. W. (2005). Ckap2 Regulates Aneuploidy, Cell Cycling, and Cell Death in a p53-Dependent Manner. *Cancer Research*, *65*(15), 6685–6691. https://doi.org/10.1158/0008-5472.CAN-04-4223

Vikas, P., Borcherding, N., & Zhang, W. (2018). The clinical promise of immunotherapy in triple-negative breast cancer. *Cancer Management and Research*, *10*, 6823–6833. https://doi.org/10.2147/CMAR.S185176

Xie, W., Zhang, J., Zhong, P., Qin, S., Zhang, H., Fan, X., Yin, Y., Liang, R., Han, Y., Liao, Y., Yu, X., Long, H., Lv, Z., Ma, C., & Yu, F. (2019). Expression and potential prognostic value of histone family gene signature in breast cancer. *Experimental and Therapeutic Medicine*, *18*(6), 4893–4903. https://doi.org/10.3892/etm.2019.8131

Yin, L., Duan, J.-J., Bian, X.-W., & Yu, S. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research*, *22*(1), 61. https://doi.org/10.1186/s13058-020-01296-5

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, L., Fang, C., Xu, X., Li, A., Cai, Q., & Long, X. (2015). Androgen receptor, EGFR, and BRCA1 as biomarkers in triple-negative breast cancer: A meta-analysis. *BioMed Research International*, *2015*, 357485. https://doi.org/10.1155/2015/357485

Acknowledgement

The research topic originated from my deep interest in cancer studies. I am fully responsible for doing the research and writing the paper. I would like to thank Ruoyao Ni (Ph.D. student) for aiding me voluntarily with his kind suggestions on the choice of the topic and the appropriate and scientific use of language in the paper.