Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning based Tissue Analysis Reveals Brachyury has a Diagnosis Value in Breast Cancer

Kaichun Li (shtumor@163.com)
Shanghai Fourth People's Hospital Affiliated to Tongji University https://orcid.org/0000-0001-6442-4388

Qiaoyun Wang
Shanghai Fourth People's Hospital Affiliated to Tongji University

Yanyan Lu
Shanghai Fourth People's Hospital Affiliated to Tongji University

Xiaorong Pan
Shanghai Fourth People's Hospital Affiliated to Tongji University

Long Liu
Tianyou Hospital Affiliated to Tongji Uniersity

Shiyu Cheng
Tianyou Hospital Affliated to Tongji University

Bingxiang Wu
Tianyou Hospital Affiliated to Tongji University

Zongchang Song
Tianyou Hospital Affiliated to Tongji University

Wei Gao
Shanghai Fourth People's Hospital Affiliated to Tongji University

Research article

Keywords: Breast cancer, Brachyury, survival

Posted Date: September 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-60846/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Bioscience Reports on April 1st, 2021. See the published version at https://doi.org/10.1042/BSR20203391.

Abstract

Background The aim of this study was to confirm the role of Brachyury in breast cells and to establish and verify whether four types of machine learning models can use Brachyury expression to predict the survival of patients.

Methods We conducted a retrospective review of the medical records to obtain patient information, and made the patient's paraffin tissue into tissue chips for staining analysis. We selected a total of 303 patients for research and implemented four machine learning prediction algorithms, including multivariate logistic regression model, decision tree, artificial neural network and random forest, and compared the results of these models with each other. Area under the receiver operating characteristic (ROC) curve (AUC) was used to compare the results.

Results The chi-square test results of relevant data suggested that the expression of Brachyury protein in cancer tissues was significantly higher than that in paracancerous tissues (p=0.0335); breast cancer patients with high Brachyury expression had a worse overall survival (OS) compared with patients with low Brachyury expression. We also found that Brachyury expression was associated with ER expression (p=0.0489). Subsequently, we used four machine learning models to verify the relationship between Brachyury expression and the survival of breast cancer patients. The results showed that the decision tree model had the best performance (AUC=0.781).

Conclusions Brachyury is highly expressed in breast cancer and indicates that the patient had a poor chance of survival. Compared with conventional statistical methods, decision tree model shows superior performance in predicting the survival status of breast cancer patients. This indicates that machine learning can thus be applied in a wide range of clinical studies.

Background

In recent years, with the improvement of computer hardware and software technology, machine learning has developed rapidly. Recently, the development of machine learning in the medical field has also received more and more attention. In the field of medical diagnosis, machine learning can clearly distinguish data noise, improve the reliability of results, and work by finding patterns in data obtained from diagnostic tests, which can be used to predict clinical outcomes or to detect obstructive phenotypes. For example, machine learning can be used to predict the response of melanoma patients to PD1 antibody treatment **[1]**. In addition, machine learning algorithms can screen out several important variables that affect antibody therapy among many factors, and build models to predict the response of patients to drugs. Similarly, researchers can use artificial intelligence technology to improve the accuracy of medical imaging diagnosis of important diseases **[2**, **3]**.

Brachyury is a T-box transcription factor, which has the function of driving EMT. Although EMT exists during the normal development of early embryonic cells, EMT in tumor cells is more active. Therefore, EMT makes tumor cells more invasive and resistant. Although in previous study, we have found that

Brachyury can promote the occurrence of EMT of breast cancer cells [4, 5], there is no clinical data supporting this. In this study, we prepared paraffin tissue from 303 cases of breast cancer tissues, constructed tissue chips, and tried to evaluate the value of Brachyury protein expression in breast cancer prognostic analysis using machine learning algorithms.

Methods

1. Clinical samples and immunohistochemistry

From 2002 to 2014, we collected paraffin specimens of cancer and paracancerous tissues of breast cancer patients from Shanghai Changhai Hospital, Shanghai Ruijin Hospital, Shanghai Xinhua Hospital and Shanghai Huangpu District Central Hospital, including 573 cases of primary breast cancer tissues and 29 cases of paracancerous normal tissues. Finally, we successfully constructed seven tissue chips, of which six were cancer tissue chips, with a total of 303 cases; one was a paracancerous normal tissue chip, with a total of 29 cases. All cases were diagnosed by comprehensive pathology and definitely confirmed as breast cancer. All patients received systemic local and/or systemic treatment including radiotherapy, surgery, chemotherapy and endocrine therapy. We obtained hospitalization number and pathology number from the medical record room, collected all original medical records corresponding to patients through the hospital internal database, collated the data of breast cancer patients, and classified the statistics according to specified indicators, including clinical characteristics, lymph node metastasis and TNM staging. We used the streptomycin avidin-peroxidase (HRP) complex method to determine the distribution of antigens in tissues and cells through the biotin streptavidin reaction. The results were judged by double-blind method. Without knowing the patient's clinical data, two experienced pathologists judged separately and reviewed the inconsistent results.

2. Scoring criteria for immunohistochemistry

For Brachyury-positive cells, the positive staining was light yellow, brownish-yellow, and brown, which were located in the nucleus. The results of immunohistochemistry were evaluated using a two-level scoring method. According to the degree of staining, positive cells ≤ 5% were judged as 0 points, 6%-25% were judged as 1 point, 26%-50% were judged as 2 points, and 51-75% were judged as 3 points, and >75% were judged as 4 points. For staining intensity, non-coloring was judged as negative and counted as 0 points, light brown was judged as weak positive (+) and counted as 1 point, dark brown was judged as strong positive (3+) and counted as 3 points, and staining between weak positive and strong positive was judged as (2+) and counted as 2 points. The comprehensive calculation was based on the product of staining intensity and percentage of positive cells, of which 0 points were judged as (-), 1-4 points were judged as (+), 5-8 points were judged as (2+) and 9-12 points were judged as (3+). A total score of 0-4 points was considered negative, and a total score of 5-12 points was considered positive.

3. Data analysis

We used the mice package in R to perform multiple imputation on missing data. First, SPSS 21.0 statistical software was used to perform univariate analysis on the data, and P<0.05 on both sides indicated that the difference was statistically significant. Then different statistical methods were used according to the specific conditions of the data. Mann-Whitney U non-parametric test was used to analyze the relationship between the expression of Brachyury protein and age, Pearson X^2 test or Fisher exact probability test was used to analyze the Brachyury expression in cancer tissues and paracancerous tissues, McNemar's test was used to analyze the Brachyury matched expression in cancer tissues and paracancerous tissues, and P<0.05 on both sides indicated that the difference was statistically significant. Subsequently, we calculated the person correlation coefficient between each variable, compared the relationship between each variable and the patient's prognosis, and then selected the variables suitable for modeling. We used logistic regression, random forest, decision tree and neural network algorithms to build clinical prediction models. All the above models were implemented using R language.

Result

Patient characteristics and immunohistochemical results

Our final tissue chips contained a total of 332 cases of breast cancer samples, including 303 cases of cancer tissues and 29 cases of paracancerous normal tissues, 28 of which were paired samples. The Brachyury protein expression was detected by IHC assay in breast cancer. Results showed that Brachyury, which was embedded in the nucleus and nuclear envelope, was overexpressed in breast cancer tissues. We conducted Pearson X^2 test on the positive expression of Brachyury in cancer tissues and paracancerous tissues. The results showed that the positive expression of Brachyury in cancer tissues was significantly higher than that in paracancerous tissues (Table 1,Figure 1). After that, we also conducted McNemar's test on the paired samples, and the results showed that the difference in the expression of Brachyury protein between cancer tissues and paracancerous tissues in the same breast cancer case was statistically significant (Table 2). Combined with our previous results, this further clarified that Brachyury protein expression might be related to the patient's prognosis. We also explored the relationship between Brachyury gene expression and patient survival in the KMPLOTTER database. The results showed that patients with high Brachyury expression had a poorer prognosis than patients with low Brachyury expression (Figure 2).

Correlation between Brachyury expression and clinical characteristics in breast cancer

The correlation between Brachyury expression and pathological parameters in breast cancer was analyzed. The results suggested that the differences between Brachyury protein expression and different ages, histological grade, tumor size, presence or absence of lymph node metastasis, AJCC stage, pathological diagnosis and PR expression status could not be considered statistically significant, and the differences between Brachyury protein expression and ER (P=0.0392) and HER2 (P=0.0572) expressions could be considered statistically significant (Table 3). Survival prognosis is one of the important basis for

clinical decision to implement specific interventions for breast cancer patients, but there is currently no recognized gold standard for prognostic analysis of breast cancer.

We used the Pearson correlation coefficient to test the correlation between various variables in breast cancer patients. The results showed that even the common pathological staging of breast cancer that frequently used in clinical practice, such as molecular typing or TNM staging, had little correlation with the survival rate of patients (Figure 3).

The performance of machine learning models

We used 75% (227 cases) of samples as the training set, and 25% (75 cases) of samples as the test set, and employed machine learning algorithms random forest, decision tree, neural network and logistic regression, all of which were superior to algorithms of conventional statistical methods, to consider Brachyury expression and other clinical variables as predictors to construct clinical predictive models for prognostic analysis of breast cancer. The results showed that the decision tree model performed best, with AUC=0.781, sensitivity=0.6 and specificity=0.894 (Figure 4A), while the other three models had AUCs less than 0.7, of which logistic regression AUC=0.665, sensitivity=0.5 and specificity=0.909 (Figure 4B); neural network AUC=0.658, sensitivity=0.4 and specificity=0.970 (Figure 4C); random forest AUC=0.645, sensitivity=0.5 and specificity=0.833 (Figure 4D). The ROC curve of decision tree model showed the highest accuracy, which indicated that it was feasible and effective to integrate the clinical variables of the patients and the pathological detection results of Brachyury as a comprehensive model for predicting the survival of breast cancer patients.

Discussion

Brachyury is one of the members of the T-box transcription factor family. Previous studies have shown that the Brachyury expression in tumors is not only related to primary tumors, but also to metastatic tumors and recurrent tumors. Our previous study has found that Brachyury in breast cancer cells can act on SIRT1 to promote Tamoxifen resistance **[6]**, indicating that Brachyury may be a therapeutic target for breast cancer. In triple negative breast cancer, Brachyury expression is also higher than normal tissues **[7]**. Brachyury can improve the invasive ability of breast cancer cells **[8]**, block the cell cycle process, and mediate the development of tumor drug resistance **[9]**. Brachyury down-regulation or knockout can increase the sensitivity of tumors to chemoradiation **[10]**, indicating that Brachyury plays an important role in the development of breast cancer. In this study, we used tissue chip technology to detect 303 postoperative breast cancer tissue samples, and the results showed that the Brachyury expression in breast cancer tissues was higher than that in paracancerous tissues. More interestingly, we found that the Brachyury expression was related to the molecular typing of breast cancer, especially the expression status of ER, which provided clinical data support for our previous point that Brachyury expression could promote patients' resistance to tamoxifen. This will encourage us to further explore the mechanism by which Brachyury causes tamoxifen resistance and evaluate its potential as a target to reverse tamoxifen resistance.

Similarly, previous studies have shown that the Brachyury expression is closely related to the prognosis of breast cancer patients. For example, the expression level of Brachyury combined with status of tumor-infiltrating CD8+ and FOXP3+ lymphocytes is used to predict the therapeutic effect of radiotherapy and chemotherapy **[11]**. Kwan Ho Lee et al. have also found that high Brachyury expression in primary breast cancer can be used as a poor prognostic factor for breast cancer **[12]**. However, although there are many prognostic indicators for breast cancer, the accuracy of single indicator is not high. The combined application of multiple indicators helps to increase the correct rate of selection and improve the treatment effect. In this study, we considered immunohistochemical staining scores of Brachyury together with the prognostic analysis indicators commonly used in the clinical practice, such as TNM staging, as a complex.

Some previous studies have clarified the advantages of machine learning in predicting disease outcomes at different sample sizes. For example, Edmond et al. developed a morphological classifier based on machine learning to distinguish different levels of epithelial dysplasia in Barrett's esophagus **[13]**. Another study used immunohistochemical results from 131 breast cancer patients to explore biomarkers of breast cancer and verified them in 65 cases of samples **[14]**. Shipp et al. also used 77 samples to predict the outcome of patients with diffuse large B-cell lymphoma **[15]**. However, a larger sample size might create a more accurate model. In this study, we used Brachyury protein staining results of tissue chips from 303 breast cancer patients, combined with relevant clinicopathological data, and applied machine learning algorithms to improve the accuracy of predicting breast cancer survival outcomes.

Our study showed that the results of decision tree model were better than conventional multivariate regression statistical models, and also better than other machine learning models .This might be due to the fact that we converted the variables into grading variables as much as possible during the research process. However, our results were not intended to indicate that we had obtained a perfect classifier. One of the major disadvantages of this study was that, although we found that Brachyury expression was related to the molecular typing of breast cancer, our limited sample size was not enough to support our use of machine learning models in different molecular typing of breast cancer to predict the impact of Brachyury staining and other pathological parameters on the survival of breast cancer patients. In subsequent studies, we plan to further collect samples of ER-positive breast cancer patients for Brachyury staining to improve our prediction model. In addition, due to the lack of intelligibility of the output of machine learning algorithm, our study does not clarify how Brachyury expression is related to the expression of ER, which also needs to be further explored in future study **[16]**.

Conclusion

We further clarified the relationship between Brachyury expression and ER in clinical samples. At the same time, we also found that one of the machine learning methods, decision tree, could effectively use Brachyury expression to predict the prognosis of breast cancer patients, and its accuracy was higher than that of conventional statistical methods.

Abbreviations

receiver operating characteristic	ROC
Area under the receiver operating characteristic curve	AUC
overall survival	OS
streptomycin avidin-peroxidase	HRP
Estrogen	ER

Declarations

Acknowledgments

Funding

This research was funded by special project of clinical research of health industry of Shanghai Municipal Health Commission(No.201940178). The funder is the first author of the study. In the study, he collected samples, participated in data analysis and drafted the manuscript.

Availability of data and materials

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

Authors' contributions

KL collected samples, participated in data analysis and drafted the manuscript. QW and LL participated in data analysis. YL carried out the immunohistochemical. XP collectted samples. SC and BW performed data analysisand contributed to the analysis of the results. ZS and WG designed the study and helped to draft the manuscript.

Competing interest

The authors declare no conflicts of interest.

Ethics approval and consent to participate

The study was approved by the ethics committee and institutional review board

of Shanghai Fourth People's Hospital Affiliated to Tongji University. The

ethics approval number is 2020031-001.

After the approval of the ethics committee, it is in compliance with the regulations that we obtained the patient's written consent.

References

1. Indini A, Di Guardo L, Cimminiello C, De Braud F, Del VM (2019) Artificial Intelligence Estimates the Importance of Baseline Factors in Predicting Response to Anti-PD1 in Metastatic Melanoma. Am J Clin Oncol 42: 643-648 Doi 10.1097/COC.0000000000000566
2. Park S, Chu LC, Fishman EK, Yuille AL, Vogelstein B, Kinzler KW, Horton KM, Hruban RH, Zinreich ES, Fouladi DFet al (2020) Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation. Diagn Interv Imaging 101: 35-44 Doi 10.1016/j.diii.2019.05.008
3. Sheth D, Giger ML (2020) Artificial intelligence in the interpretation of breast cancer on MRI. J Magn Reson Imaging 51: 1310-1324 Doi 10.1002/jmri.26878
4. Li K, Ying M, Feng D, Du J, Chen S, Dan B, Wang C, Wang Y (2016) Brachyury promotes tamoxifen resistance in breast cancer by targeting SIRT1. Biomed Pharmacother 84: 28-33 Doi 10.1016/j.biopha.2016.09.011
5. Li K, Ying M, Feng D, Chen Y, Wang J, Wang Y (2016) SMC1 promotes epithelial-mesenchymal transition in triple-negative breast cancer through upregulating Brachyury. Oncol Rep 35: 2405-2412 Doi 10.3892/or.2016.4564
6. Li K, Ying M, Feng D, Du J, Chen S, Dan B, Wang C, Wang Y (2016) Brachyury promotes tamoxifen resistance in breast cancer by targeting SIRT1. Biomed Pharmacother 84: 28-33 Doi 10.1016/j.biopha.2016.09.011
7. Hamilton DH, Roselli M, Ferroni P, Costarelli L, Cavaliere F, Taffuri M, Palena C, Guadagni F (2016) Brachyury, a vaccine target, is overexpressed in triple-negative breast cancer. Endocr Relat Cancer 23: 783-796 Doi 10.1530/ERC-16-0037
8. Pires MM, Aaronson SA (2014) Brachyury: a new player in promoting breast cancer aggressiveness. J Natl Cancer Inst 106 Doi 10.1093/jnci/dju094
9. Huang B, Cohen JR, Fernando RI, Hamilton DH, Litzinger MT, Hodge JW, Palena C (2013) The embryonic transcription factor Brachyury blocks cell cycle progression and mediates tumor resistance to conventional antitumor therapies. Cell Death Dis 4: e682 Doi 10.1038/cddis.2013.208
10. Kobayashi Y, Sugiura T, Imajyo I, Shimoda M, Ishii K, Akimoto N, Yoshihama N, Mori Y (2014) Knockdown of the T-box transcription factor Brachyury increases sensitivity of adenoid cystic carcinoma cells to chemotherapy and radiation in vitro: implications for a new therapeutic principle. Int J Oncol 44: 1107-1117 Doi 10.3892/ijo.2014.2292
11. Lee KH, Kim EY, Park YL, Do SI, Chae SW, Park CH (2017) Expression of epithelial-mesenchymal transition driver brachyury and status of tumor-infiltrating CD8+ and FOXP3+ lymphocytes in predicting treatment responses to neoadjuvant chemotherapy of breast cancer. Tumour Biol 39: 1393379089 Doi 10.1177/1010428317710575

12. Lee KH, Kim EY, Yun JS, Park YL, Do SI, Chae SW, Park CH (2018) Prognostic significance of expression of epithelial-mesenchymal transition driver brachyury in breast cancer and its association with subtype and characteristics. Oncol Lett 15: 1037-1045 Doi 10.3892/ol.2017.7402
13. Sabo E, Beck AH, Montgomery EA, Bhattacharya B, Meitner P, Wang JY, Resnick MB (2006) Computerized morphometry as an aid in determining the grade of dysplasia and progression to adenocarcinoma in Barrett's esophagus. Lab Invest 86: 1261-1271 Doi 10.1038/labinvest.3700481
14. Popovici V, Budinská E, Čápková L, Schwarz D, Dušek L, Feit J, Jaggi R (2016) Joint analysis of histopathology image features and gene expression in breast cancer. Bmc Bioinformatics 17: 209 Doi 10.1186/s12859-016-1072-z
15. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GSet al (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8: 68-74 Doi 10.1038/nm0102-68
16. McDonald L, Ramagopalan SV, Cox AP, Oguz M (2017) Unintended consequences of machine learning in medicine? F1000Res 6: 1707 Doi 10.12688/f1000research.12693.1

Tables

Table I. Expression of Brachyury protein in breast cancer and paracancerous tissues.

	N	Negative	Positive	X_square	p_value
Tumor	303	209(68.98)	94(31.02)	4.5181	0.0335
Paracancerous	29	26(89.66)	3(10.34)		

Table II. Expression of Brachyury protein in paired cases of breast cancer and paracancerous tissues.

	Negative	Positive	N(%)	X_square	p_value
Paracancerous Negative	12(42.86)	13(46.43)	25(89.29)	8.6429	0.0033
Paracancerous Positive	1(3.57)	2(7.14)	3(10.71)		
N(%)	13(46.43)	15(53.57)	28(100)		

Table III. Relationship between Brachyury protein expression and clinical pathological parameters of breast cancer.

		Negative	Positive	χ^2	P
Age	Median age	53(30-83)	54(30-84)		0.1302
	AJCC stage				
stage:1	I	43(63.24)	25(36.76)	1.9009	0.3866
stage:2	II	112(69.14)	50(30.86)		
stage:3	III	54(73.97)	19(26.03)		
	Histological stage				
hyphology_class_new_y:1	I	4(66.67)	2(33.33)	0.1568	0.9246
hyphology_class_new_y:2	II	147(69.67)	64(30.33)		
hyphology_class_new_y:3	III	58(67.44)	28(32.56)		
	Menstrual status				
menopause:0	Menopause	111(65.29)	59(34.71)	2.0783	0.1494
menopause:1	Not menopausal	98(73.68)	35(26.32)		
	Tumor size				
Tumour_max_diameter:<=2cm	≤2cm	75(65.79)	39(34.21)	3.2908	0.1929
Tumour_max_diameter:2.1~5cm	2.1–5cm	116(69.05)	52(30.95)		
Tumour_max_diameter:>5cm	>5cm	18(85.71)	3(14.29)		
	Lymph node metastasis				
lymph_node:0	0	118(69.41)	52(30.59)		
lymph_node:1~3	1–3	45(67.16)	22(32.84)		
lymph_node:4~9	4–9	27(77.14)	8(22.86)		
lymph_node:>=10	≥10	19(61.29)	12(38.71)	2.0645	0.5591
	Molecular type				
molecular.type:1	Luminal A	48(60.76)	31(39.24)	8.1353	0.0433
molecular.type:2	Luminal B	7(58.33)	5(41.67)		
molecular.type:3	Her2 overexpression	31(86.11)	5(13.89)		
molecular.type:4	Triple negative	123(69.89)	53(30.11)		
	ER				

ER_value_new_y:1	-	154(72.64)	58(27.36)	3.8781	0.0489
ER_value_new_y:2	+	55(60.44)	36(39.56)		
	PR				
PR_value_new_y:1	-	184(70.5)	77(29.5)	1.5556	0.2123
PR_value_new_y:2	+	25(59.52)	17(40.48)		
	Her2				
HER2:-	-	137(65.55)	72(34.45)	3.1985	0.0737
HER2:+	+	72(76.6)	22(23.4)		

Figures

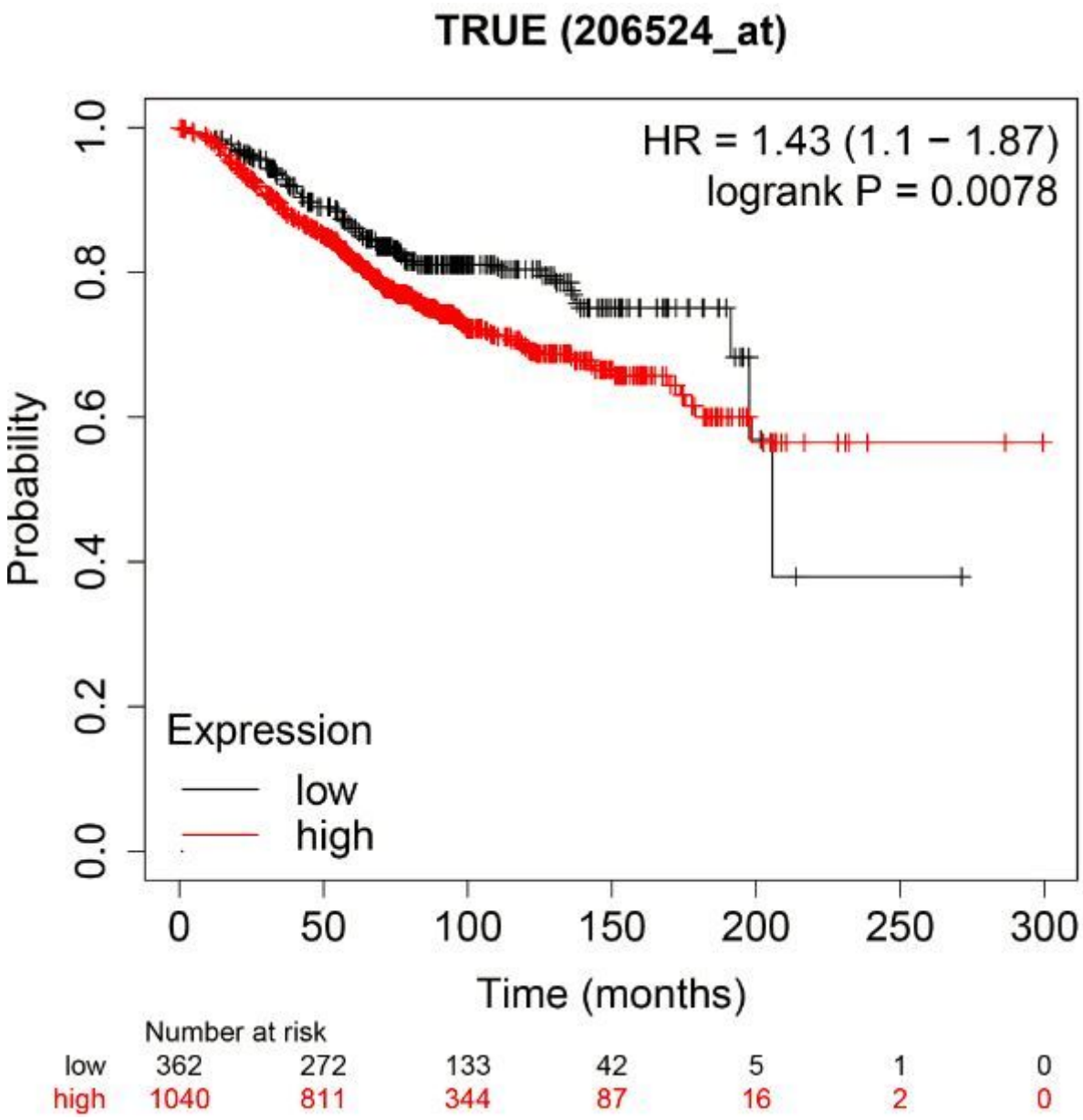

Figure 1

Expression of Brachyury protein in breast cancer tissues (A. cancer tissue +, B. cancer tissue ++, C. cancer tissue +++, D. paracancerous tissue +)

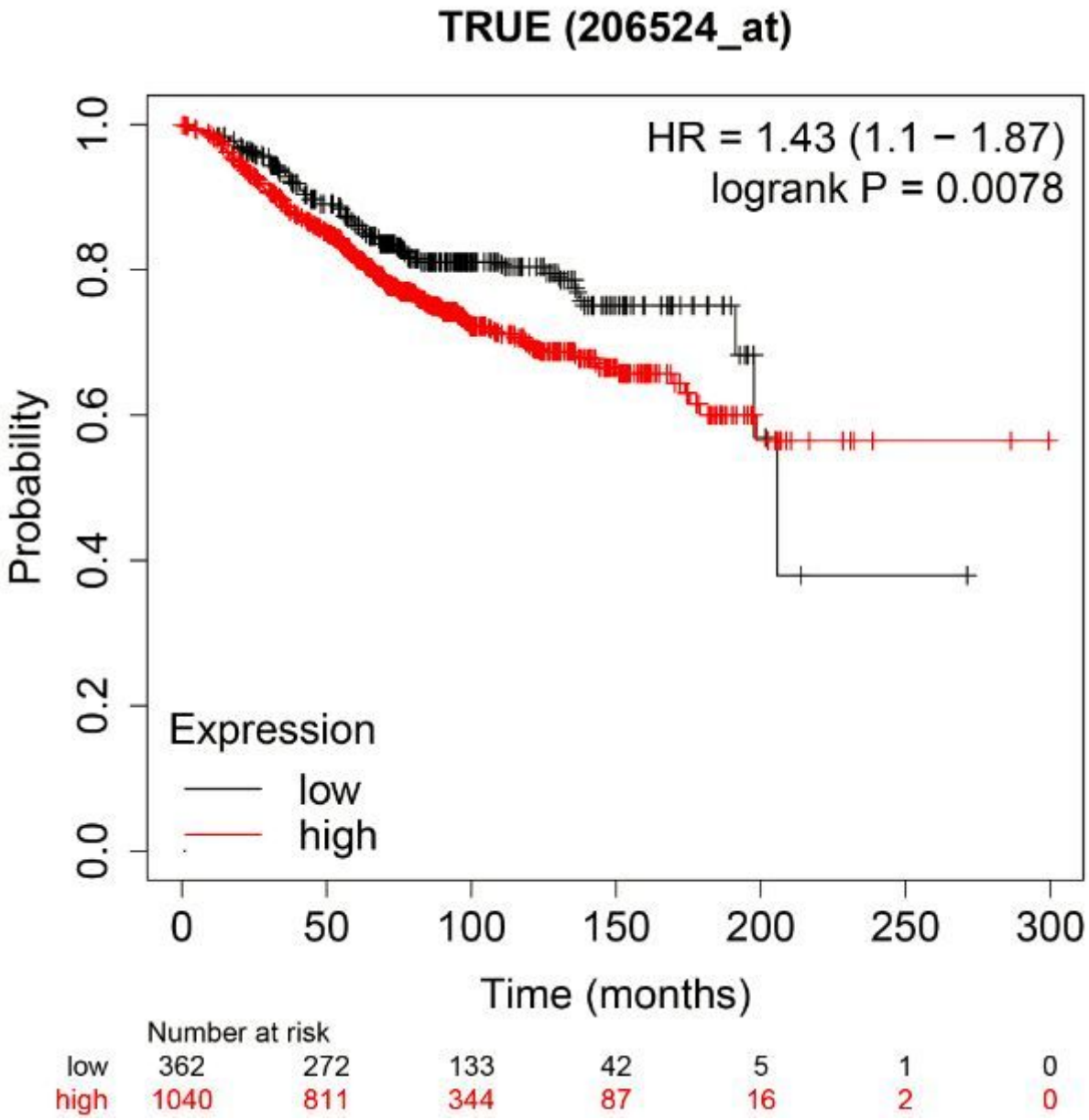

Figure 2

Survival values of Brachyury expression generated by the Kaplan-Meier (KM) plotter.

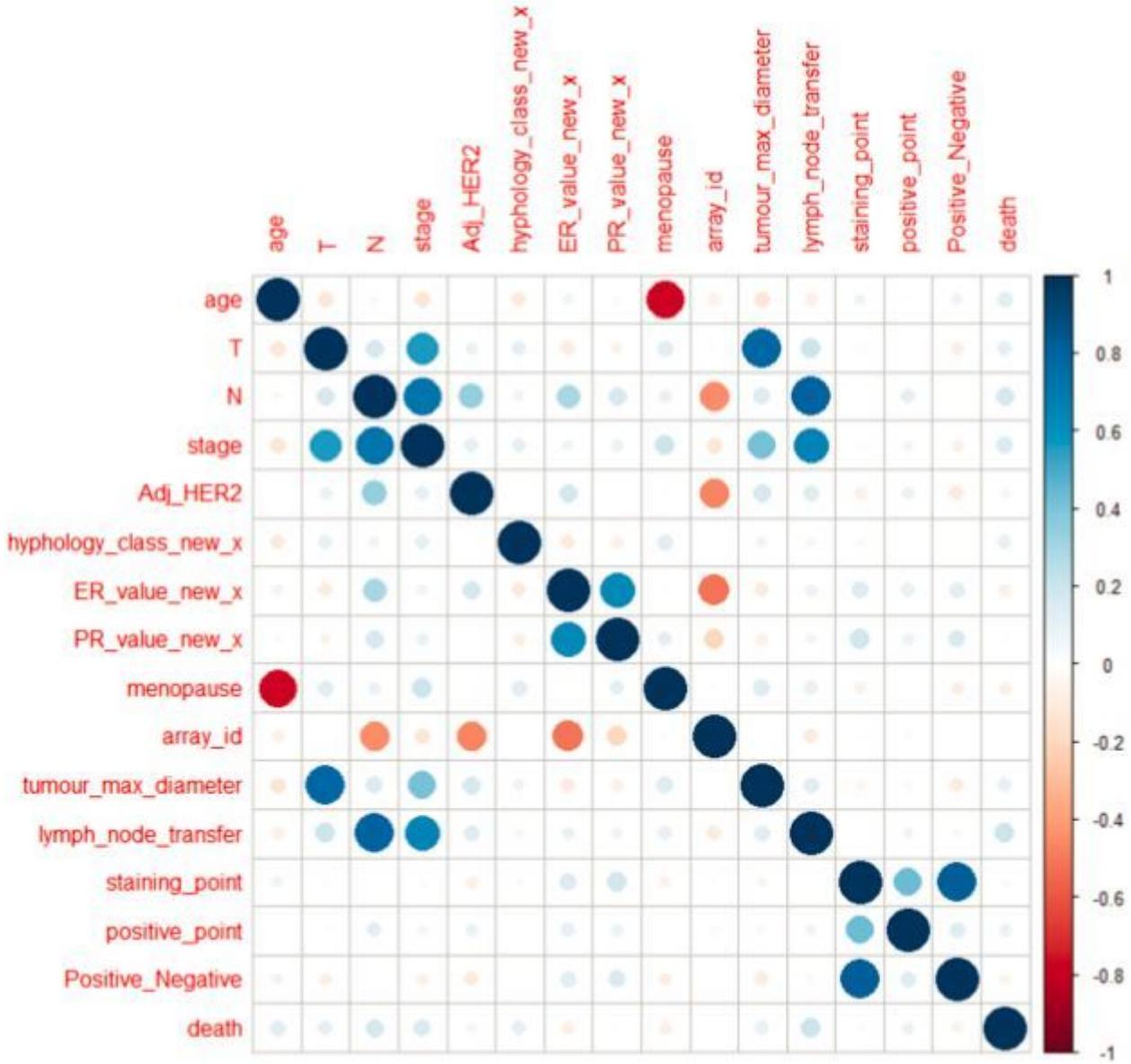

Figure 3

Pearson correlation matrix of breast cancer patient data.

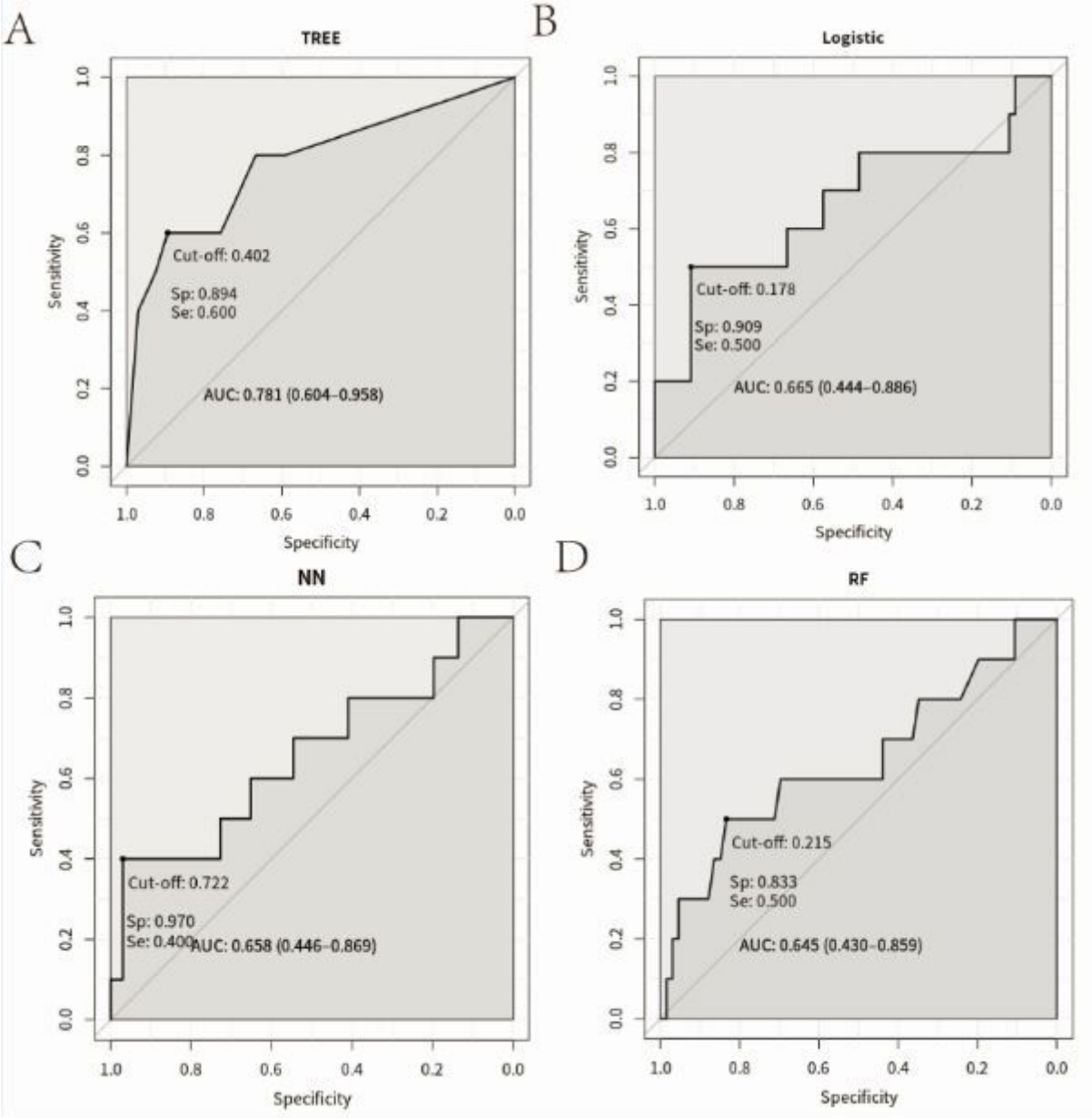

Figure 4

ROC curves used to assess model performance. A. Decision tree B: Logistic regression C. Neural network D. Random forest