

Received: 2020.11.23

Accepted: 2021.02.13

Available online: 2021.02.28

Published: 2021.05.18

Identification of a DNA Methylation-Based Prognostic Signature for Patients with Triple-Negative Breast Cancer

Authors' Contribution:

Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

ABCE 1 **Yinqi Gao**
AC 2 **Xuelong Wang**
CE 1 **Shihui Li**
CE 1 **Zhiqiang Zhang**
CE 1 **Xuefei Li**
AE 1 **Fangcai Lin**

1 Department of Breast Surgery, Capital Medical University Electric Power Teaching Hospital, Beijing, China

2 Department of Thoracic Surgery, Capital Medical University Electric Power Teaching Hospital, Beijing, P.R. China

Corresponding Author: Fangcai Lin, e-mail: linfangcai0704@126.com
Source of support: Departmental sources

Background: Aberrant DNA methylation is an important biological regulatory mechanism in malignant tumors. However, it remains underutilized for establishing prognostic models for triple-negative breast cancer (TNBC).


Material/Methods: Methylation data and expression data downloaded from The Cancer Genome Atlas (TCGA) were used to identify differentially methylated sites (DMSs). The prognosis-related DMSs were selected by univariate Cox regression analysis. Functional enrichment was analyzed using DAVID. A protein-protein interaction (PPI) network was constructed using STRING. Finally, a methylation-based prognostic signature was constructed using LASSO method and further validated in 2 validation cohorts.

Results: Firstly, we identified 743 DMSs corresponding to 332 genes, including 357 hypermethylated sites and 386 hypomethylated sites. Furthermore, we selected 103 prognosis-related DMSs by univariate Cox regression. Using a LASSO algorithm, we established a 5-DMSs prognostic signature in TCGA-TNBC cohort, which could classify TNBC patients with significant survival difference (log-rank $p=4.97E-03$). Patients in the high-risk group had shorter overall survival than patients in the low-risk group. The excellent performance was validated in GSE78754 (HR=2.42, 95%CI: 1.27-4.59, log-rank $P=0.0055$). Moreover, for disease-free survival, the prognostic performance was verified in GSE141441 (HR=2.09, 95%CI: 1.28-3.44, log-rank $P=0.0027$). Multivariate Cox regression analysis indicated that the 5-DMSs signature could serve as an independent risk factor.

Conclusions: We constructed a 5-DMSs signature with excellent performance for the prediction of disease-free survival and overall survival, providing a guide for clinicians in directing personalized therapeutic regimen selection of TNBC patients.

Keywords: **DNA Methylation • Prognosis • Triple Negative Breast Neoplasms**

Full-text PDF: <https://www.medscimonit.com/abstract/index/idArt/930025>

 3653

 5

 7

 44



Background

Breast cancer (BC) is a common malignant tumor that occurs in the thymic epithelial tissue, and it has become the main cause of mortality and morbidity in women [1,2]. Triple-negative breast cancer (TNBC), defined as the loss of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor-2 (HER2) expression, is a particularly aggressive subtype of BC, accounting for 15-20% of all breast cancers [3]. Compared to other types of BC, patients with TNBC usually have poorer overall survival (OS) and a higher probability of cancer recurrence due to the lack of early detection biomarkers and clear therapeutic targets [4]. Hence, it is urgent to develop robust prognostic biomarkers to guide personalized therapy.

DNA methylation is one of the most commonly occurring epigenetic events that regulate gene expression without altering the DNA sequence [5]. Generally, cytosine residues in the CpGs, frequently found in the proximal promoter regions of many genes, are methylated by DNA methyltransferases that catalyze the transfer of the active methyl group from S-adenosylmethionine to the C5-position of cytosine [6-8]. It has been shown that hypermethylation at promoter CpG islands typically results in decreased transcription of downstream genes [9]. When methylation is experimentally removed from promoter regions, transcription levels rise [10]. Increasing studies have revealed that epigenetic gene silencing caused by DNA methylation could play important roles in tumorigenesis and progression [11-13]. It has also been found that the aberrant DNA methylation could be an important risk factor in TNBC [14,15]. For example, Prajzencanc et al found that BRCA1 promoter methylation in peripheral blood cells was frequently detected in TNBC using methylation-sensitive high-resolution melting, which was significantly associated with reduced BRCA1 expression, aggressive phenotype, and poor prognosis [16]. BRCA1 hypermethylation confers a homologous recombination deficiency, immune cell type, genome-wide DNA methylation, and transcriptional phenotype similar to TNBC tumors with BRCA1-inactivating variants. Given that DNA methylation changes are plausibly critical components of the molecular mechanisms involved in TNBC, distinct DNA methylation could be a potential biomarker to improve the accuracy of TNBC diagnosis and prognosis. In addition, the emergence of high-throughput technology makes it possible to identify reliable biomarkers for diagnosis and prognosis. Several studies have proposed many DNA methylation signatures to improve the accuracy of cancer prognosis [17-19]. However, few robust DNA methylation biomarkers were identified as available indicators. Although a nomogram incorporating a prognostic risk model and clinicopathological features was developed by Peng et al for predicting the prognosis of TNBC [20], more reliable methylation-based biomarkers to predict the survival are needed for personalized therapy of TNBC patients.

In the present study, based on the DNA methylation data downloaded from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), we identified differentially methylated sites (DMSs) with consistent difference. Combining gene expression data and clinical data of TNBC samples downloaded from TCGA, we further screened the prognosis-related DMSs with aberrant gene expression. Based on prognosis-related DMSs, we aimed to establish a prognostic signature to classify TNBC patients into different risk groups. The prognostic performance was validated in 2 independent cohorts. Such a signature could provide effective help for prognosis classification and personalized therapy of TNBC patients.

Material and Methods

Data Collection

TNBC samples and normal breast samples with methylation data (Platform: Illumina Infinium HumanMethylation450), mRNA-Seq data (Platform: Illumina HiSeq 2000 RNA sequencing), and clinical information were downloaded from TCGA (<https://gdc-portal.nci.nih.gov/>). Patients with unknown survival time, age, tumor stage, and lymph node metastasis status were excluded. Ultimately, 114 TNBC samples and 86 normal breast samples were retained in our study. A DNA methylation cohort (Platform: Illumina Infinium HumanMethylation450) was derived from the GEO database, which contains 14 TNBC samples and 17 normal breast samples. Moreover, 2 DNA methylation cohorts (Platform: Illumina Infinium HumanMethylation450) with clinical information (GSE141441 and GSE78754), containing 130 and 63 TNBC samples, respectively, were downloaded from the GEO database as the validation cohorts. All cohorts analyzed in the present study are included in **Table 1**. Approval by the Institutional Ethics Committee was not necessary because all data were collected from the publicly available GEO and TCGA databases.

Screening of Differentially Methylated Sites

The DNA methylation data, defined as β values, were derived from the GEO and TCGA databases. The methylation data processed by background subtraction, quantile normalization, and quality control had normal distribution. Thus, the t test was used to selected differentially methylated sites (DMSs) between TNBC and adjacent normal breast tissues. Sites with FDR <0.05 and consistent difference direction in both GEO and TCGA cohorts were selected as DMSs. Moreover, comparing the expression data of TNBC and adjacent normal breast tissues from TCGA, we identified differentially expressed genes (DEGs) using the edgeR package with FDR <0.05 and $|\log_2FC| > 2$ as the threshold. The differentially methylated probes were annotated by the platform annotation files of Illumina

Table 1. Cohorts analyzed in the present study.

	GSE52865		TCGA-TNBC		GSE141441		GSE78754	
	Methylation data	Methylation data	Expression data	Clinic data	Methylation data	Clinic data	Methylation data	Clinic data
Normal	17	86	86	–	–	–	–	–
Tumor	14	114	114	114	130	130	63	63
Platform	Illumina HM450	Illumina HM450	Illumina HiSeqV2		Illumina HM450		Illumina HM450	

HumanMethylation450 Bead-Chip. Correlation between DMSs and corresponding DEGs was calculated by Pearson correlation coefficient. DMSs with $P < 0.05$ and coefficient < -0.15 were screened for downstream analysis.

Functional Enrichment Analysis

Function enrichment, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, was performed for DEGs corresponding to DMSs using the Database for Annotation, Visualization and Integrated Discovery (DAVID, <https://david.ncifcrf.gov/>) to investigate the functions and biological pathways regulated by DMSs. GO enrichment includes biology process (BP), cellular component (CC), and molecular function (MF). KEGG provides insight into biological pathways. The significant terms and pathways were selected with $P < 0.05$ as the threshold.

Protein–Protein Interaction (PPI) Network Analysis

The PPI network provides a meaningful insight into the molecular mechanisms of key cellular activities in cancer progression. Based on DEGs containing DMSs, a PPI network was established using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (<https://string-db.org/>). A score for each interaction between DEGs, ranging from 0 to 1, was calculated by STRING. A higher score means higher reliability. An interaction score of 0.7 was regarded as the cut-off criterion to select crucial PPI interaction pairs. Cytoscape (version 3.7.1; www.cytoscape.org) was used to visualize the PPI network. Using the cytoHubba application, hub genes with high degrees of connectivity were identified.

Construction of Prognostic Signature

Using univariate Cox regression analysis, DMSs significantly associated with overall survival (OS) were identified with a $P < 0.05$ threshold in TCGA-TNBC cohort, which were defined as prognosis-related DMSs. To better investigate the performance of those DMSs in predicting prognosis, a prognostic model was built by LASSO method using the glmnet R package. Removing

the prognosis-related DMSs with coefficient equal to 0, an optimal prognostic scoring model was subsequently constructed with the following formula:

$$\text{Risk score} = \sum \text{Coef}_{\text{site}} \times \text{Methylation}_{\text{site}}$$

$\text{Coef}_{\text{site}}$ denote the coefficient of a prognosis-related DMS. $\text{Methylation}_{\text{site}}$ is the methylation level of the same prognosis-related DMS. A risk score was calculated for each patient using the prognostic model. Then, patients were classified into a high-risk group and a low-risk group based on the median risk score.

Statistical Analysis

The t test was used to select DMSs with $\text{FDR} < 0.05$. The binomial test was performed to observe the difference in consistency of DMSs. DEGs were screened using the edgeR package with $\text{FDR} < 0.05$ and $|\log_2\text{FC}| > 2$ as the threshold. Pearson correlation coefficient analysis was performed to calculate the correlation between DMSs and the corresponding DEGs. Clinical factors, including age, stage, and lymph node metastasis status, were selected and categorized to perform Cox regression analysis. Age was categorized with a 60-year threshold. Stage I and II were categorized as low-stage, while stage III and IV were categorized as high-stage. We used the Cox regression model for estimating hazard ratios (HRs) and 95% confidence intervals (CIs). Multivariate Cox regression analysis was performed to evaluate the independent prognostic performance of the methylation-based signature after adjusting for clinical factors. As they are not recorded in GSE141441 and GSE78754, clinical factors such as stage and lymph node metastasis status were excluded when performing Cox regression analysis. Kaplan-Meier plots were used to illustrate survival differences between different risk groups using the log-rank test. The mortality rate and lymph node metastasis rate of high-risk and low-risk groups were compared using Fisher's exact test. A P value < 0.05 was considered to indicate a statistically significant difference. All statistical analyses were performed using R3.4.0.

Table 2. Clinical characteristics of the patients in the training and validation cohorts.

Variables	Training cohort TCGA-TNBC N(%)	Validation Cohort GSE141441 N(%)	Validation Cohort GSE78754 N(%)
Sample			
Normal	–	–	–
Tumor	114 (100%)	130 (100%)	63 (100%)
Mean age (years; range)	55 (29-84)	57 (32-92)	54 (30-83)
Stage			
I	19 (16.7%)	44 (33.8%)	–
II	73 (64.0%)	72 (55.4%)	–
III	20 (17.5%)	14 (10.8%)	–
IV	2 (1.8%)	0 (0.0%)	–
Lymph node status			
Metastasis	71 (62.3%)	–	–
Non-metastasis	43 (37.7%)	–	–
Vital status			
Alive	100 (87.7%)	–	20 (31.7%)
Dead	14 (12.3%)	–	43 (68.3%)
Relapse status			
Yes	–	68 (52.3%)	–
No	–	62 (47.7%)	–

Results

Patient Characteristics

A total of 114 patients collected from TCGA were included as the training cohort for constructing the prognostic signature. In the training cohort, most patients had stage II disease (64.7%), and only 2 patients had stage IV disease. Seventy-one (62.3%) patients had lymph node metastasis. In the validation cohort GSE141441, most patients also had stage II disease (55.4%), and no patients with stage IV were included. Furthermore, more than half of patients had recurrence. In the validation cohort GSE78754, tumor stage was not recorded and the vital status of most patients was death. The clinicopathological characteristics of the patients are summarized in **Table 2**.

Identification of Differentially Methylated Sites in TNBC

With cutoff criteria of FDR <0.05, 250 024 and 180 148 DMSs were selected from TCGA-TNBC and GSE52865 cohorts, respectively. As shown in **Figure 1A**, a total of 110 228 DMSs, including

49 199 hypermethylation sites and 61 029 hypomethylation sites, were selected with 95.9% concordance score (binomial test, $P < 0.0001$). Moreover, we identified 1049 DEGs with cutoff criteria of FDR <0.05 and $|\log_{2}FC| > 2$ (**Figure 1B**), including 568 upregulated genes and 481 downregulated genes. To identify DMSs that directly regulate gene expression, we calculated the Pearson correlation coefficient to observe the association between the methylation value of a DMS and the expression value of the corresponding DEG. A total of 743 DMSs, whose methylation levels were significantly negatively associated with expression levels of 332 corresponding DEGs, were selected for further analysis.

Functional Enrichment

Functional enrichment analysis was performed using DAVID. The significant terms enriched from each GO category are shown in **Figure 2A**. The results showed that several cancer-related terms were significantly enriched. The most significant GO terms were angiogenesis ($P < 0.001$). Extracellular matrix organization, inflammatory response, and cell adhesion were also

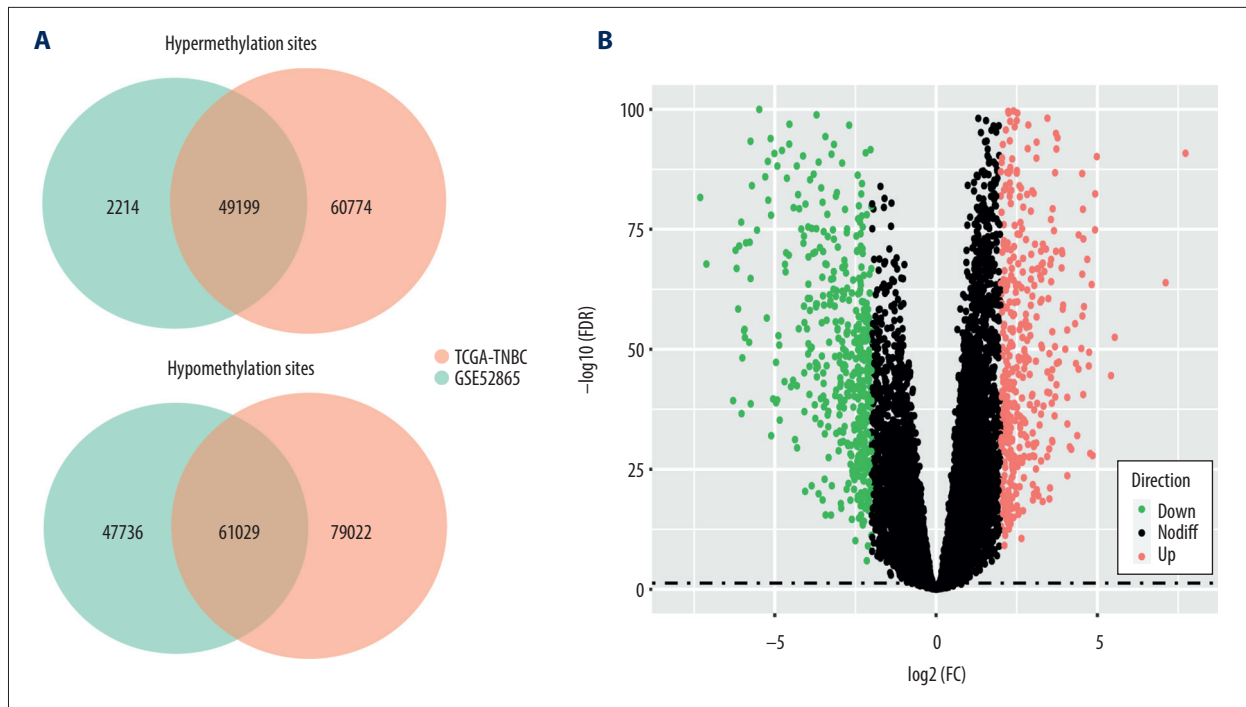


Figure 1. Consistency of differentially methylated sites and identification of differentially expressed genes. **(A)** Overlapping differentially methylated sites between TCGA-TNBC and GSE52865. **(B)** Volcano plot of differentially expressed genes.

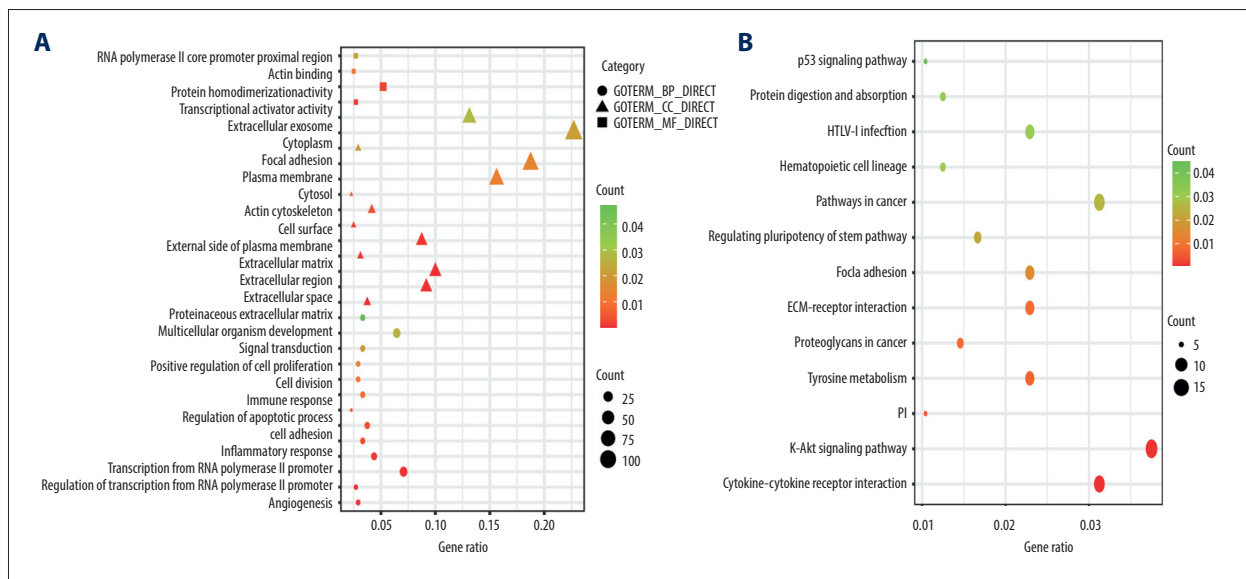


Figure 2. Functional enrichment of the significantly enriched Gene Ontology terms **(A)** and Kyoto Encyclopedia of Genes and Genomes pathways **(B)** of differentially expressed genes containing differentially methylated sites.

significantly enriched. As shown in **Figure 2B**, KEGG analysis suggested that the DEGs were significantly enriched in cytokine-cytokine receptor interaction, PI3K-Akt signaling pathway, and ECM-receptor interaction. Functional enrichment analysis indicated that DMSs affected the occurrence and progression of TNBC by regulating the corresponding gene expression.

Construction of PPI Network

To further explore the interactions between the 332 DEGs, a PPI network was constructed. With an interaction score >0.7 as the cutoff criterion, a PPI network containing 168 nodes and 396 edges was constructed using STRING (**Figure 3A**). Furthermore, as shown in **Figure 3B**, 10 hub genes with high

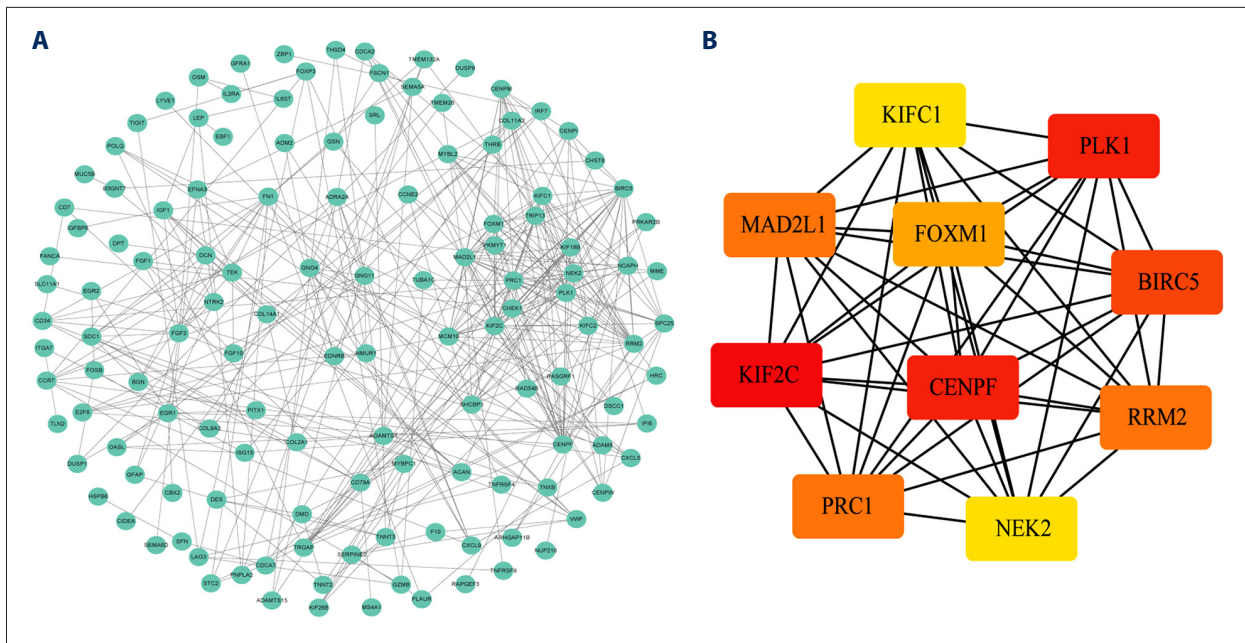


Figure 3. Identification of hub genes within protein-protein interaction (PPI) network. (A) PPI network. (B) Top 10 hub genes with high degrees of connectivity.

Table 3. General information of the 5 differentially methylated sites.

ProbeID	Gene	Chrom	ChromStart	ChromEnd	Relation to CpG island	Relation to gene region
cg15724876	TGFBR2	Chr3	30714670	30714672	Open sea	Body
cg17887364	EIF4EBP1	Chr8	37891956	37891958	Shelf	Body
cg19419246	FOSB	Chr19	45950424	45950426	Open sea	Intergenic
cg21234506	BCL2A1	Chr15	80263131	80263133	Island	Promoter
cg21580376	ADRB2	Chr5	148206411	148206413	Island	Promoter

degrees of connectivity within the PPI network were identified using the cytoHubba plugin for Cytoscape, including kinesin family member 2C (KIF2C), polo-like kinase 1 (PLK1), centromere protein F (CENPF), baculoviral IAP repeat containing 5 (BIRC5), mitotic arrest deficient 2-like 1 (MAD2L1), protein regulator of cytokinesis 1 (PRC1), ribonucleotide reductase regulatory subunit M2 (RRM2), forkhead box M1 (FOXM1), NIMA-related kinase 2 (NEK2), and kinesin family member C1 (KIFC1).

Establishment of a 5-DMSs Prognostic Signature

A total of 103 prognosis-related DMSs, including 46 hypermethylation sites and 57 hypomethylation sites, were identified by univariate Cox proportional hazards regression analysis ($P < 0.05$). **Supplementary Table 1** contains a list of 103 DMSs. Based on these prognosis-related DMSs, we established a 5-DMSs signature to guide the prediction

of prognosis for TNBC patients using the LASSO algorithm. The features of 5-DMSs are shown in **Table 3**. Thus, the prognostic scoring model was constructed as follows: Risk score= $(-1.664 \times$ methylation degree of cg15724876) $+(-0.373 \times$ methylation degree of cg17887364) $+(-0.249 \times$ methylation degree of cg19419246) $+(0.753 \times$ methylation degree of cg21234506) $+(0.149 \times$ methylation degree of cg21580376). Two of the 5-DMSs were related to high risk (cg21234506 and cg21580376; HR >1), while 3 of the 5-DMSs appeared to be protective (cg15724876, cg17887364, and cg19419246; HR <1). We calculated the risk scores for patients in TCGA-TNBC cohort and then classified them into high-risk and low-risk groups using the median risk score as the cutoff. We found that patients in the high-risk group had poorer OS than those in the low-risk group (log-rank $P=4.97E-03$, **Figure 4A**). Multivariate Cox regression analysis showed that the 5-DMSs signature could serve as an independent prognostic factor (HR=7.55, 95% CI=1.39-40.9,

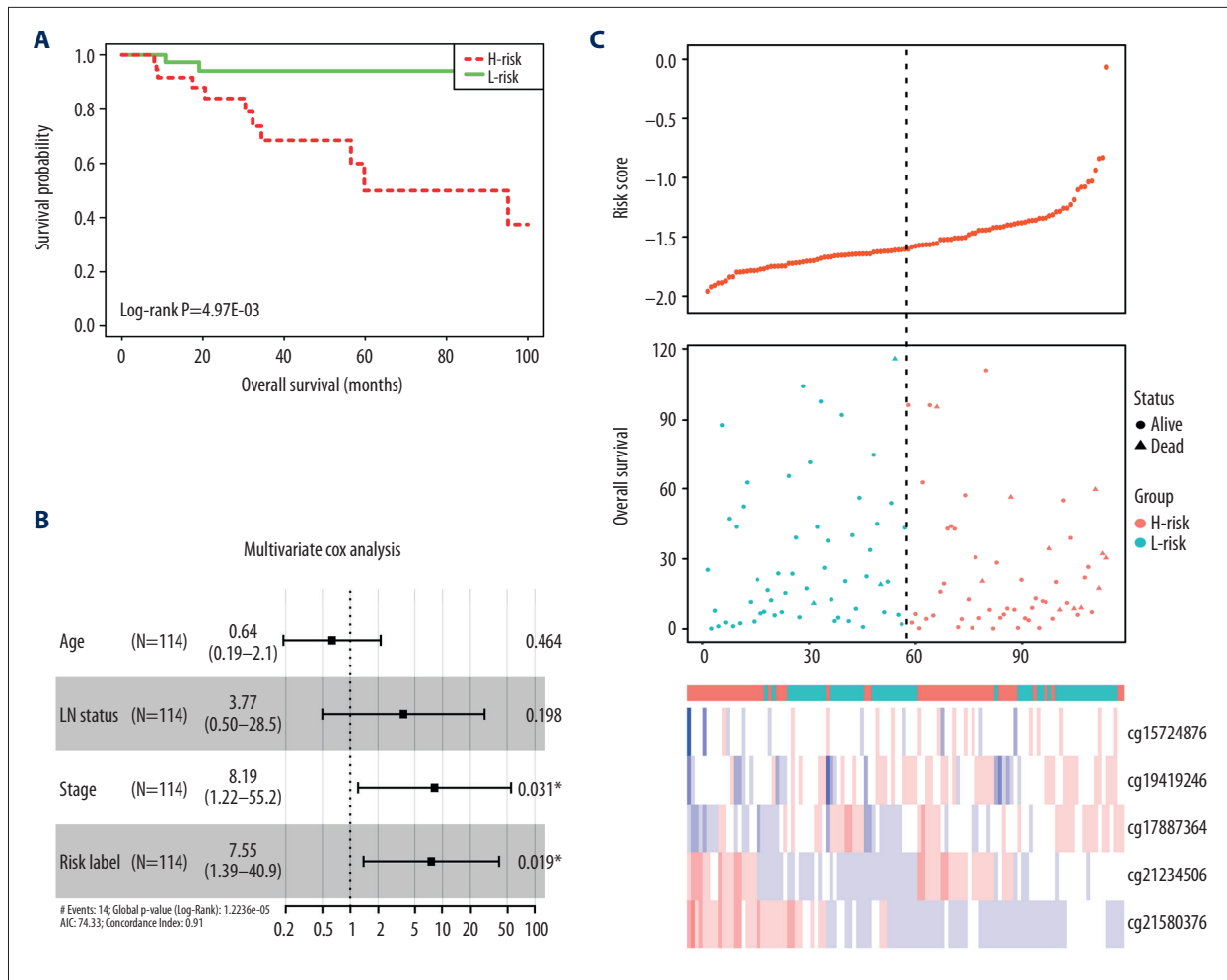


Figure 4. Evaluating the prognostic power of the 5-DMSs signature in the TCGA-TNBC cohort. **(A)** Kaplan-Meier survival curve of the high- and low-risk groups. **(B)** Multivariate analysis of risk factors for TNBC. **(C)** Distribution of risk score, survival status, and methylation heatmap of the 5-DMSs.

$P=0.019$, **Figure 4B**). The distributions of the risk scores, survival time, survival status, and 5-DMSs methylation profiles are shown in **Figure 4C**. We also performed the stratification analysis for stage to prove an independent prediction of the 5-DMSs prognostic signature. For 92 patients with stage I or II, 47 patients were classified into the low-risk group, and the rest were classified into the high-risk group. Kaplan-Meier survival curve analysis showed that patients in the high-risk group had significantly shorter OS (log-rank $P=0.015$, **Figure 5A**). Of the 22 patients with stage III or IV, 10 patients were classified into the low-risk group, and the rest were classified into the high-risk group. Although there was no statistically significant survival difference between the 2 risk groups, it still had a trend of poor prognosis in the high-risk group (log-rank $P=0.248$, **Figure 5B**). The methylation differences of the 5-DMSs between high-risk and low-risk groups are shown in **Figure 6A**. The results showed that patients in the low-risk group had significantly higher methylation degrees of 3 protective DMSs,

while patients in the high-risk group had significantly higher methylation degrees of 1 risky DMS. Although another risky DMS was not statistically significant, it still had a trend of high methylation degree in the high-risk group.

Performance Validation of the 5-DMSs Prognostic Signature

Using the 5-DMSs prognostic signature, each patient in GSE78754 was classified into a high-risk ($n=32$) or a low-risk ($n=31$) group. Kaplan-Meier survival curve analysis showed that high-risk scores were significantly correlated with poor prognosis (log-rank $P=0.0055$, **Figure 7B**). Multivariate Cox regression analysis also confirmed the independent prognostic performance of the 5-DMSs signature (HR=2.66, 95%CI=1.38-5.11, $P=0.003$, **Table 4**). Similar to findings in the training cohort, the methylation differences of each DMS were observed (**Figure 6C**). Moreover, another independent cohort (GSE141441) containing

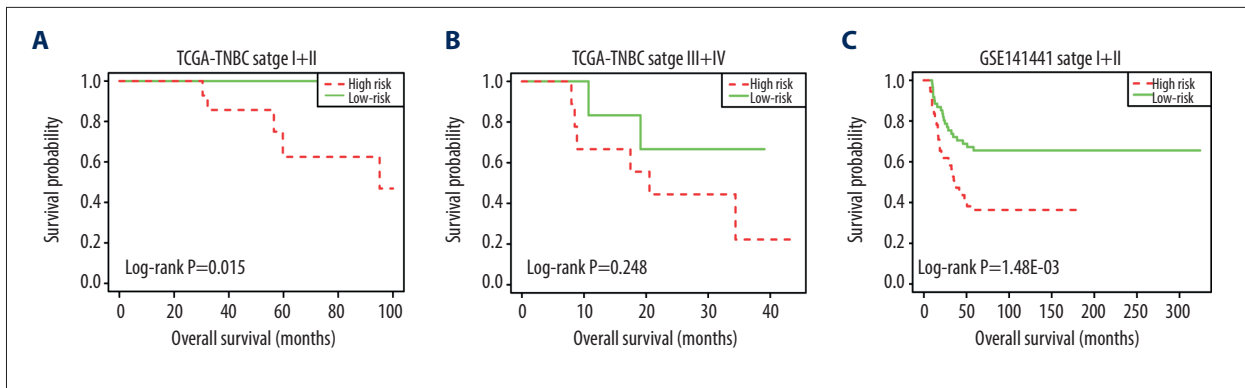


Figure 5. The stratification analysis for stage in TCGA-TNBC (A) and GSE141441 (B).

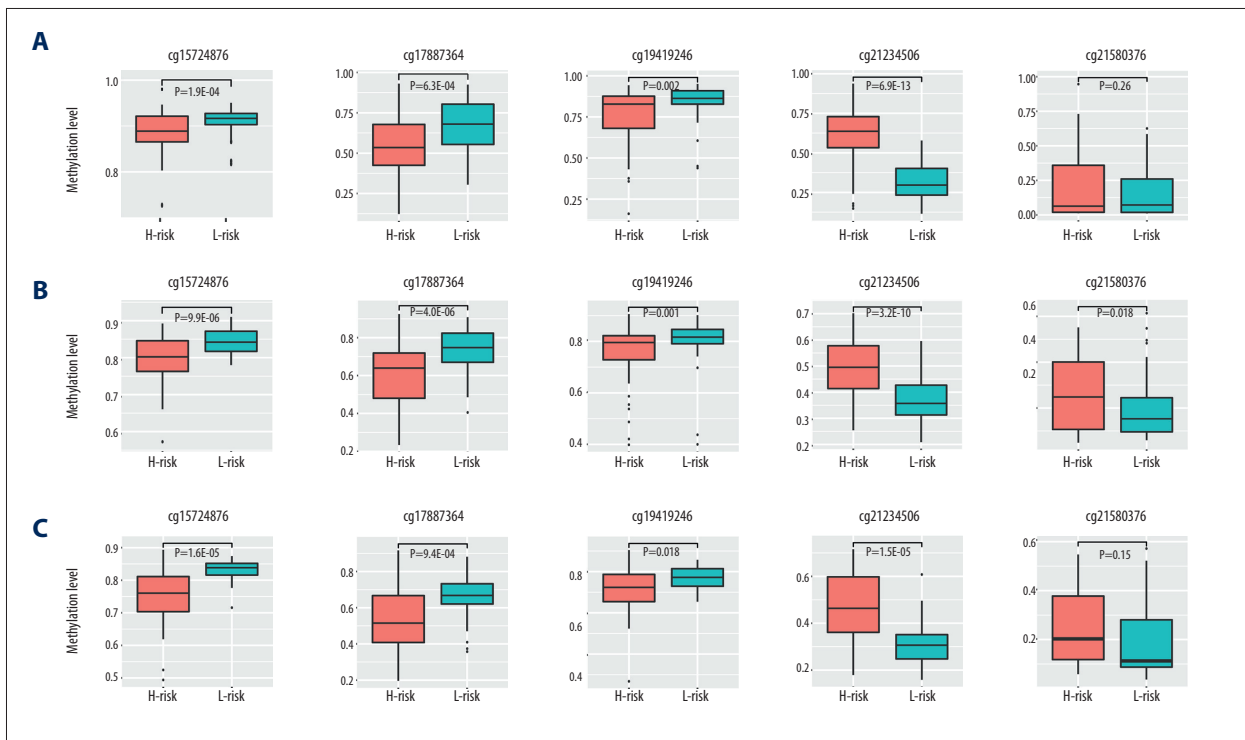


Figure 6. Methylation difference analysis of 5 differentially methylated sites within different risk groups in TCGA-TNBC (A), GSE141441 (B), and GSE78754 (C).

130 patients with disease-free survival data was used to validate the prognostic performance for tumor recurrence. The results showed that patients in the low-risk group had a longer disease-free survival (log-rank $P=0.0027$, **Figure 7A**). Multivariate Cox regression analysis suggested that the 5-DMs signature could serve as an independent prognostic factor for disease-free survival (HR=1.77, 95%CI=1.05-3.00, $P=0.033$, **Table 4**). The stratification analysis for stage also proved an independent prediction of the 5-DMs prognostic signature. For 116 patients with stage I or II, 65 patients were classified into the low-risk group and the rest were classified into the high-risk group. Kaplan-Meier survival curve analysis showed that patients in the high-risk group had significantly shorter

survival than those in the low-risk group (log-rank $p=1.48E-03$, **Figure 5C**). Moreover, all patients with stage III or IV were classified into the high-risk group. Observing the methylation differences of each DMS, we found that patients in the high-risk group tended to have higher methylation degrees at 2 risky sites, whereas patients in the low-risk group tended to have higher methylation degrees at 3 protective sites (**Figure 6B**). The risk scores distributions, survival time, survival status, and 5-DMs methylation profiles for patients in GSE141441 and GSE78754 cohorts are shown in **Figure 7C and 7D**, respectively.

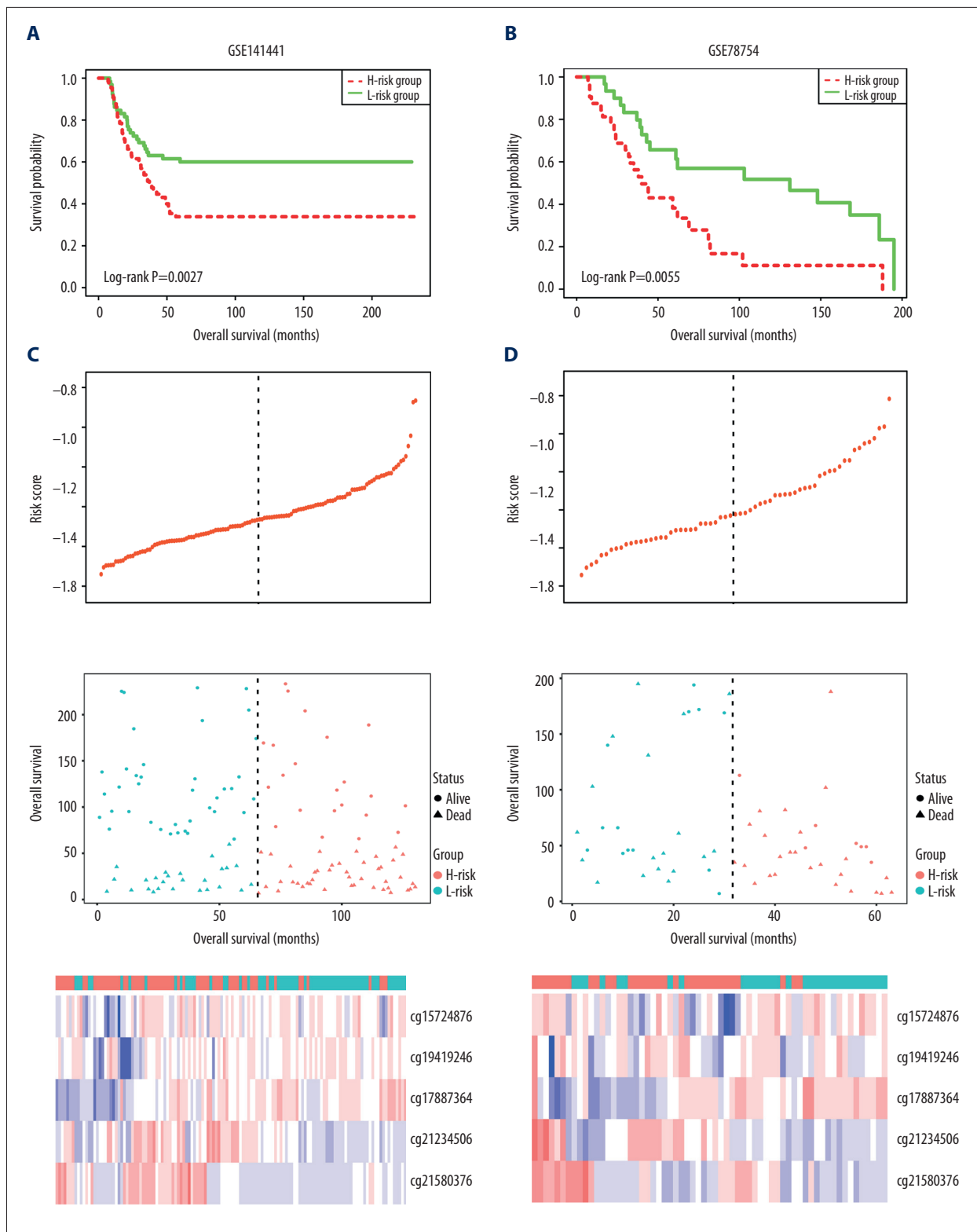


Figure 7. Kaplan-Meier survival curve in GSE141441 (A) and GSE78754 (B), and distribution of risk score, survival status, and methylation heatmap of 5 DMSs in GSE141441 (C) and GSE78754 (D).

Table 4. Univariate and multivariate Cox regression analysis in the training and validation cohorts.

Variables	Univariate analysis		Multivariate analysis	
	HR (95% CI)	P	HR (95% CI)	P
TCGA-TNBC cohort				
Age				
≤60/>60	1.02 (0.31-3.33)	0.970	0.64 (0.19-2.14)	0.464
Stage				
I-II/III+IV	19.7 (4.17-93.4)	1.7E-04	8.19 (1.22-55.2)	0.031
Lymph node status				
Non-metastasis/metastasis	4.52 (1.24-16.5)	0.022	3.77 (0.50-28.5)	0.198
Risk score				
Low/high	6.54 (1.44-29.6)	0.015	7.55 (1.39-40.9)	0.019
GSE141441 cohort				
Age				
≤60/>60	1.33 (0.82-2.14)	0.245	1.24 (0.76-2.01)	0.394
Stage				
I-II/III+IV	2.65 (1.41-4.97)	0.002	2.11 (1.09-4.09)	0.026
Risk score				
Low/high	2.09 (1.28-3.44)	0.003	1.77 (1.05-3.00)	0.033
GSE78754 cohort				
Age				
≤60/>60	0.85 (0.45-1.62)	0.616	0.66 (0.34-1.29)	0.227
Risk score				
Low/high	2.42 (1.27-4.59)	0.007	2.66 (1.38-5.11)	0.003

Discussion

TNBC is the most fatal subtype of breast cancer, which is a heterogeneous group of tumors with one common feature: a distinctly aggressive nature with higher rates of relapse and shorter overall survival in the metastatic setting [21,22]. DNA methylation refers to heritable and modifiable markers that regulate gene expression without changing the DNA sequence, which could be a potential biomarker to improve the accuracy of TNBC prognosis. In the present study, which comprehensively analyzed methylation data, expression data, and clinical data from TCGA, we identified prognosis-related DMSs and further developed a novel 5-DMSs prognostic signature, which had improved ability to predict prognosis for TNBC patients compared with clinical risk factors. Multivariate Cox regression analysis showed that the 5-DMSs signature could serve as an independent prognostic factor to categorize TNBC

patients with different outcomes. The prognostic performance for overall survival and disease-free survival was validated in independent cohorts. Our study provides a reliable prognostic signature for TNBC patients based on DNA methylation, and also identified potential therapeutic targets.

Functional enrichment analysis provided an intuitive overview of the mechanism of TNBC. The most significant GO terms were angiogenesis, a core hallmark of advanced cancer. Many studies have demonstrated that angiogenesis is important in the transformation of hyperplastic in situ epithelium to invasive carcinoma for TNBC growth and metastasis [23-25]. Other significantly enriched GO terms, such as extracellular matrix organization, inflammatory response, and cell adhesion, are also closely associated with TNBC metastasis and chemoresistance [26-28]. KEGG pathway analysis showed significant enrichment of cytokine-cytokine receptor interaction and

PI3K-Akt signaling pathway. Cytokine-cytokine receptor interaction is an important way to regulate cell growth and proliferation and participate in immune response and inflammation, especially in tumor growth and metastasis [29-31]. The PI3K-Akt signaling pathway is an important and active pathway involved in chemoresistance and survival, which is speculated to play an important part in malignant transformation and has been considered as a potential molecular target for the design of therapeutic agents to treat TNBC [32-34]. These findings indicate that the identified DMSs affect the occurrence and progression of TNBC by regulating the corresponding gene expression.

We constructed a PPI network to investigate the functional interactions between DEGs and identified 10 hub genes with the highest degree of connectivity from the PPI network. Some of these genes have been reported to be closely associated with TNBC, such as CENPF, PLK1, and KIF2C. Sun et al found that CENPF could promote BC bone metastasis by activating PI3K-AKT-mTORC1 signaling and might be a novel therapeutic target for BC treatment [35]. Ueda et al found that PLK1 was significantly overexpressed in TNBC tissues and plays a pivotal role in the regulation of mitosis of TNBC cells, which could be an attractive molecular-targeted therapy for TNBC [36].

In our study, based on LASSO regression analysis, we developed a novel prognostic signature based on 5 methylation sites (cg15724876, cg17887364, cg19419246, cg21234506, and cg21580376) corresponding to 5 genes: transforming growth factor beta receptor 2 (TGFBR2), eukaryotic translation initiation factor 4E binding protein 1 (EIF4EBP1), FBJ murine osteosarcoma viral oncogene homolog B (FOSB), BCL2-related protein A1 (BCL2A1), and adrenoceptor beta 2 (ADRB2). Cox regression analysis showed that 3 of 5 sites (cg15724876, cg17887364, and cg19419246) were protective for TNBC, and their methylation degrees were significantly higher in the low-risk group classified by prognostic signature. Moreover, the other 2 sites were significantly hypermethylated in the high-risk group as risky factors. Although few studies have reported the biological mechanism of 5 methylation sites in tumor progression, the epigenetic changes of several corresponding genes have been reported to be related to carcinogenesis, migration, and invasion. Zhang et al reported that EZH2-mediated histone 3 trimethylated on lysine 27 (H3K27me3), a marker of silent chromatin conformation, at the FOSB promoter, inhibited its expression. EZH2 inhibitor promotes the shift from H3K27me3 to H3K27ac at the FOSB promoter, and recruits the transcription factor C/EBP β to activate FOSB gene transcription. Epigenetic inactivation of FOSB mediated by EZH2 increased the gene expression in TNBC samples and further promoted TNBC cells proliferation, indicating a critical role of FOSB methylation in the regulation of TNBC progression [37]. Ma et al reported that TGFBR2, an important tumor suppressor mediating

TGF- β signaling and inducing cell cycle arrest, is downregulated in esophageal squamous cell carcinoma due to DNA hypermethylation of its promoter regions, and is involved in malignant transformation of esophageal squamous cell carcinoma by inducing cell cycle G2/M arrest [38]. All 5 genes have been reported to be involved in many cancer-related pathways to regulate TNBC progression [39-43]. For instance, Karlsson et al found that EIF4EBP1 and S6K2 have a correlated mRNA expression, and that high levels of EIF4EBP1 and S6K2 were associated with a poor prognosis, independent of other classical prognostic markers [44]. In our study, we found that EIF4EBP1 at cg17887364 was hypomethylated in TNBC, which might regulate the gene overexpression and further induce poor prognosis. However, no previous studies have reported on the association between abnormal methylation of TGFBR2, EIF4EBP1, BCL2A1, or ADRB2 and TNBC. Hence, it is necessary to perform further studies on the methylation regulation mechanism of these genes and TNBC.

Cancer is a complex regulatory network in which multisite methylation as biomarkers could achieve higher specificity and sensitivity compared with single-site methylation. Peng et al [20] developed a 15-CpG-based signature for predicting prognosis in TNBC based on a TCGA-TNBC cohort to make a prognosis classification for TNBC patients. Unfortunately, the insufficient external validation limited it as a clinically reliable indicator. Our 5-DMSs signature was carefully developed with a particular focus on this issue. In our study, using the combination of weighted methylation values of these 5-DMSs, we constructed a 5-DMSs signature for the prediction of prognosis and validated the prognostic performance with a larger sample size containing 2 external cohorts. In methodology, we further evaluated the association between methylation levels of DMSs and expression levels of the corresponding DEGs, allowing us to screen for more meaningful DMSs related to tumor progression. However, limitations still exist in this study. First, although this study was validated using GEO data, prospective studies and multicenter clinical trials are needed to further clinical validation for the 5-DMSs prognostic signature. Second, lack of mechanism studies hinders better application of the 5-DMSs prognostic signature. Therefore, experimental research on each DMS within the 5-DMSs prognostic signature should be performed, which might provide significant information to further the understanding of their functional roles.

Conclusions

In summary, comprehensively analyzing the methylation data, expression data, and clinical data, we constructed a potential prognostic tool, named the 5-DMSs signature, to predict prognosis for patients with TNBC, which might provide support for clinicians in directing personalized therapeutic regimen selection.

Availability of Data and Materials

The data used and analyzed during the current study are available from The Cancer Genome Atlas (TCGA) (<https://cancergenome.nih.gov/>) and Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>).

Conflict of Interest

None.

Supplementary Data

Supplementary Table 1. List of prognosis-related differentially methylated sites.

Probe	Coef	HR	P.value	Gene symbol
cg00311768	-4.36342	0.012735	0.047414	TSTA3
cg00411595	9.499339	13350.9	0.041674	SLC2A6
cg00666842	-2.90891	0.054535	0.016011	SMYD1
cg00699831	24.28987	3.54E+10	0.015735	SCD
cg01119512	-11.4891	1.02E-05	0.044574	GRHL3
cg01414194	-4.13226	0.016047	0.022339	TSTA3
cg01820374	5.957623	386.6897	0.048866	LAG3
cg02187231	8.532535	5077.302	0.015247	VAX2
cg02222728	3.95713	52.30698	0.039254	SOX17
cg02311932	4.740023	114.4368	0.034187	RBM24
cg02369775	12.83755	376076.6	0.033092	IGFBP6
cg02462195	3.25937	26.03313	0.029795	ADAMTS5
cg03199006	45.4402	5.43E+19	0.002709	CXCL9
cg03323067	4.310195	74.455	0.034005	TNS1
cg03709663	-3.91068	0.020027	0.029031	GINS2
cg04370730	-4.40706	0.012191	0.034052	GJB3
cg04454576	-5.04614	0.006434	0.009928	ASCL2
cg04456029	13.33355	617569.6	0.047443	DTX1
cg04833514	-4.03234	0.017733	0.013399	SLC6A9
cg05008975	-3.4041	0.033237	0.047412	SCN4B
cg05526099	19.15269	2.08E+08	0.030173	CHI3L1
cg05814654	3.507149	33.35305	0.042772	IL21R
cg06195379	4.665694	106.2393	0.046127	IL32
cg07210669	-4.59045	0.010148	0.012612	S100P
cg07712493	2.601348	13.48189	0.027394	CDO1
cg07747553	-5.13931	0.005862	0.049336	RASD2
cg07763231	-4.41222	0.012128	0.021462	SLC16A3
cg07991704	-3.3905	0.033692	0.008325	SLC4A4

Supplementary Table 1 continued. List of prognosis-related differentially methylated sites.

Probe	Coef	HR	P.value	Gene symbol
cg08142094	-4.46022	0.01156	0.031394	GIN52
cg08480266	-3.1528	0.042732	0.043551	LGI4
cg08816023	-3.38711	0.033806	0.03887	FGF1
cg09183450	-5.50468	0.004068	0.010343	AHNAK
cg09284708	-5.51481	0.004027	0.013559	MOCS1
cg09340386	-2.32479	0.097804	0.041455	TPO
cg09757109	-3.36292	0.034634	0.037446	DIXDC1
cg10140583	-3.92209	0.0198	0.008001	LY6D
cg10241809	-6.36823	0.001715	0.045466	SLC16A3
cg10270306	3.962011	52.56294	0.042718	AKAP12
cg10725316	2.871826	17.66924	0.034809	KCNIP2
cg10846969	64.18582	7.51E+27	0.014242	CENPW
cg10988368	-5.0549	0.006378	0.005625	ASCL2
cg11348249	6.264056	525.3456	0.032164	AKAP12
cg11580525	-3.5706	0.028139	0.045714	SCN4B
cg11644479	-3.08738	0.045621	0.035826	ASCL2
cg12574296	-3.58437	0.027754	0.0409	NCCRP1
cg12989650	22.12875	4.08E+09	0.036625	ARHGEF15
cg13551227	3.624474	37.50498	0.033813	KCNIP2
cg13681655	-4.50336	0.011072	0.036481	SDC1
cg13936863	7.885756	2659.134	0.015535	VAX2
cg13971504	12.29604	218828.7	0.012028	GCNT4
cg14042396	-5.03589	0.0065	0.013656	MYCN
cg14171514	-3.69861	0.024758	0.032232	AHNAK
cg14235768	5.698774	298.5012	0.048681	VAX2
cg14588178	-3.53185	0.029251	0.039197	GDF10
cg15108410	-3.83748	0.021548	0.019046	COL2A1
cg15171154	-6.35385	0.00174	0.047176	TGFBR2
cg15724876	-4.24938	0.014273	0.005053	TGFBR2
cg16350494	26.97487	5.19E+11	0.02856	FOXP3
cg17306740	2.586213	13.27938	0.047608	ZBP1
cg17520539	5.614495	274.3749	0.028113	AKAP12
cg17863139	-5.8516	0.002875	0.048414	LAD1
cg17887364	-3.66561	0.025589	0.008412	EIF4EBP1
cg17967059	4.153925	63.6835	0.04683	GJB2

Supplementary Table 1 continued. List of prognosis-related differentially methylated sites.

Probe	Coef	HR	P.value	Gene symbol
cg18183642	-4.05077	0.017409	0.014428	SLC6A9
cg18250832	-7.23907	0.000718	0.022473	NMUR1
cg18438461	-2.22567	0.107995	0.044718	TPO
cg18567954	13.83816	1022913	0.027343	DTX1
cg18825531	-4.72894	0.008836	0.042292	AHNAK
cg18967533	-4.366	0.012702	0.019672	KLK6
cg19419246	-4.16463	0.015535	0.004675	FOSB
cg19421218	7.610693	2019.678	0.028452	TIGIT
cg19570321	-3.52385	0.029486	0.020144	HRASLS5
cg19699289	16.14231	10245148	0.040794	VAX2
cg19914554	7.227266	1376.454	0.049081	CD7
cg20000220	-3.47643	0.030918	0.043647	MAPK15
cg20062691	-4.74399	0.008704	0.009959	ISG15
cg20392764	-4.67557	0.00932	0.008544	ASCL2
cg20518446	-1.98147	0.137866	0.048057	AHNAK
cg20914659	-3.89854	0.020271	0.018332	KCNK5
cg20989454	-4.30745	0.013468	0.032761	IRF7
cg21083175	-3.88221	0.020605	0.036995	RCOR2
cg21166775	-5.21559	0.005431	0.016962	RAPGEF3
cg21214613	-6.40936	0.001646	0.019865	HSPB7
cg21234506	4.270848	71.58232	0.004655	BCL2A1
cg21580376	2.54752	12.77539	0.009966	ADRB2
cg22077313	-6.30806	0.001822	0.019035	S100P
cg22833618	7.06031	1164.806	0.0212	HRC
cg22946658	-4.78056	0.008391	0.00482	ASCL2
cg23492779	-4.97062	0.006939	0.012658	PI16
cg23568324	-4.31416	0.013378	0.040389	TTYH3
cg23588217	3.814983	45.37598	0.036747	TBX15
cg24617203	-3.28597	0.037404	0.01673	IL1R2
cg25433648	-4.1992	0.015008	0.038807	S100A14
cg25714865	-2.59033	0.074995	0.010923	FSCN1
cg25752754	-4.47079	0.011438	0.00175	SDC1
cg26051413	-2.60156	0.074158	0.03539	ASCL2
cg26584545	-3.83307	0.021643	0.017717	NTN4
cg27146050	-5.99987	0.002479	0.040453	HIF3A

Supplementary Table 1 continued. List of prognosis-related differentially methylated sites.

Probe	Coef	HR	P.value	Gene symbol
cg27245646	-2.73164	0.065113	0.032031	PODXL2
cg27347104	19.93238	4.53E+08	0.043137	VWF
cg27409154	-4.05648	0.01731	0.043397	IGFBP6
cg27450924	-3.54134	0.028974	0.023418	ARL9
ch.1.4190055F	10.84239	51143.43	0.028872	CENPF

Coef – regression coefficients; HR – hazard ratio.

References:

- Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin*. 2018;68(6):394-424
- Harbeck N, Gnant M. Breast cancer. *Lancet*. 2017;389(10074):1134-50
- Dent R, Trudeau M, Pritchard KI, et al. Triple-negative breast cancer: Clinical features and patterns of recurrence. *Clin Cancer Res*. 2007;13(15 Pt 1):4429-34
- Qiu J, Xue X, Hu C, et al. Comparison of clinicopathological features and prognosis in triple-negative and non-triple negative breast cancer. *J Cancer*. 2016;7(2):167-73
- Rana AK, Ankri S. Reviving the RNA world: An insight into the appearance of RNA methyltransferases. *Front Genet*. 2016;7:99
- Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33(Suppl.):245-54
- Serman A, Vlahovic M, Serman L, Bulic-Jakus F. DNA methylation as a regulatory mechanism for gene expression in mammals. *Coll Antropol*. 2006;30(3):665-71
- Weber M, Hellmann I, Stadler MB, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*. 2007;39(4):457-66
- Stein R, Razin A, Cedar H. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc Natl Acad Sci USA*. 1982;79(11):3418-22
- Hansen RS, Gartler SM. 5-Azacytidine-induced reactivation of the human X chromosome-linked PGK1 gene is associated with a large region of cytosine demethylation in the 5' CpG island. *Proc Natl Acad Sci USA*. 1990;87(11):4174-78
- Akhavan-Niaki H, Samadani AA. DNA methylation and cancer development: Molecular mechanism. *Cell Biochem Biophys*. 2013;67(2):501-13
- Dehan P, Kustermans G, Guenin S, et al. DNA methylation and cancer diagnosis: new methods and applications. *Expert Rev Mol Diagn*. 2009;9(7):651-57
- Klutstein M, Nejman D, Greenfield R, Cedar H. DNA methylation in cancer and aging. *Cancer Res*. 2016;76(12):3446-50
- Nakai K, Xia W, Liao HW, et al. The role of PRMT1 in EGFR methylation and signaling in MDA-MB-468 triple-negative breast cancer cells. *Breast Cancer*. 2018;25(1):74-80
- Xu J, Sun T, Guo X, et al. Estrogen receptor-alpha promoter methylation is a biomarker for outcome prediction of cisplatin resistance in triple-negative breast cancer. *Oncol Lett*. 2018;15(3):2855-62
- Prajzandanc K, Domagala P, Hybiak J, et al. BRCA1 promoter methylation in peripheral blood is associated with the risk of triple-negative breast cancer. *Int J Cancer*. 2020;146(5):1293-98
- Li C, Zheng Y, Pu K, et al. A four-DNA methylation signature as a novel prognostic biomarker for survival of patients with gastric cancer. *Cancer Cell Int*. 2020;20:88
- Tao C, Luo R, Song J, et al. A seven-DNA methylation signature as a novel prognostic biomarker in breast cancer. *J Cell Biochem*. 2020;121(3):2385-93
- Wang Y, Wang Y, Wang Y, Zhang Y. Identification of prognostic signature of non-small cell lung cancer based on TCGA methylation data. *Sci Rep*. 2020;10(1):8575
- Peng Y, Shui L, Xie J, Liu S. Development and validation of a novel 15-CpG-based signature for predicting prognosis in triple-negative breast cancer. *J Cell Mol Med*. 2020 Aug;24(16):9378-87
- Carey LA, Dees EC, Sawyer L, et al. The triple negative paradox: Primary tumor chemosensitivity of breast cancer subtypes. *Clin Cancer Res*. 2007;13(8):2329-34
- Liedtke C, Mazouni C, Hess KR, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol*. 2008;26(8):1275-81
- Braicu C, Chiorean R, Irimie A, et al. Novel insight into triple-negative breast cancers, the emerging role of angiogenesis, and antiangiogenic therapy. *Expert Rev Mol Med*. 2016;18:e18
- Ribatti D, Nico B, Ruggieri S, et al. Angiogenesis and antiangiogenesis in triple-negative breast cancer. *Transl Oncol*. 2016;9(5):453-57
- Wang C, Li J, Ye S, et al. Oestrogen inhibits VEGF expression and angiogenesis in triple-negative breast cancer by activating GPER-1. *J Cancer*. 2018;9(20):3802-11
- Matsumoto H, Koo SL, Dent R, et al. Role of inflammatory infiltrates in triple negative breast cancer. *J Clin Pathol*. 2015;68(7):506-10
- Nakazawa Y, Taniyama Y, Sanada F, et al. Periostin blockade overcomes chemoresistance via restricting the expansion of mesenchymal tumor subpopulations in breast cancer. *Sci Rep*. 2018;8(1):4013
- Nekulova M, Holcakova J, Gu X, et al. DeltaNp63alpha expression induces loss of cell adhesion in triple-negative breast cancer cells. *BMC Cancer*. 2016;16(1):782
- Barbie TU, Alexe G, Aref AR, et al. Targeting an IKBKE cytokine network impairs triple-negative breast cancer growth. *J Clin Invest*. 2014;124(12):5411-23
- Li M, Wang Y, Wei F, An X, et al. Efficiency of cytokine-induced killer cells in combination with chemotherapy for triple-negative breast cancer. *J Breast Cancer*. 2018;21(2):150-57
- Qiao Y, He H, Jonsson P, et al. AP-1 is a key regulator of proinflammatory cytokine TNFalpha-mediated triple-negative breast cancer progression. *J Biol Chem*. 2016;291(10):5068-79
- Khan MA, Jain VK, Rizwanullah M, et al. PI3K/AKT/mTOR pathway inhibitors in triple-negative breast cancer: A review on drug discovery and future challenges. *Drug Discov Today*. 2019;24(11):2181-91
- Costa RLB, Han HS, Gradishar WJ. Targeting the PI3K/AKT/mTOR pathway in triple-negative breast cancer: A review. *Breast Cancer Res Treat*. 2018;169(3):397-406
- Deng F, Weng Y, Li X, et al. Overexpression of IL-8 promotes cell migration via PI3K-Akt signaling pathway and EMT in triple-negative breast cancer. *Pathol Res Pract*. 2020;216(4):152902
- Sun J, Huang J, Lan J, et al. Overexpression of CENPF correlates with poor prognosis and tumor bone metastasis in breast cancer. *Cancer Cell Int*. 2019;19:264
- Ueda A, Oikawa K, Fujita K, et al. Therapeutic potential of PLK1 inhibition in triple-negative breast cancer. *Lab Invest*. 2019;99(9):1275-86

37. Zhang R, Li X, Liu Z, et al. EZH2 inhibitors-mediated epigenetic reactivation of FOSB inhibits triple-negative breast cancer progress. *Cancer Cell Int.* 2020;20:175
38. Ma Y, He S, Gao A, et al. Methylation silencing of TGF-beta receptor type II is involved in malignant transformation of esophageal squamous cell carcinoma. *Clin Epigenetics.* 2020;12(1):25
39. Xie F, Jin K, Shao L, et al. FAF1 phosphorylation by AKT accumulates TGF-beta type II receptor and drives breast cancer metastasis. *Nat Commun.* 2017;8:15021
40. Coleman LJ, Peter MB, Teall TJ, et al. Combined analysis of eIF4E and 4E-binding protein expression predicts breast cancer survival and estimates eIF4E activity. *Br J Cancer.* 2009;100(9):1393-99
41. Hiraki M, Maeda T, Mehrotra N, et al. Targeting MUC1-C suppresses BCL2A1 in triple-negative breast cancer. *Signal Transduct Target Ther.* 2018;3:13
42. Kurozumi S, Kaira K, Matsumoto H, et al. beta2-Adrenergic receptor expression is associated with biomarkers of tumor immunity and predicts poor prognosis in estrogen receptor-negative breast cancer. *Breast Cancer Res Treat.* 2019;177(3):603-10
43. Ting CH, Chen YC, Wu CJ, et al. Targeting FOSB with a cationic antimicrobial peptide, TP4, for treatment of triple-negative breast cancer. *Oncotarget.* 2016;7(26):40329-47
44. Karlsson E, Perez-Tenorio G, Amin R, et al. The mTOR effectors 4EBP1 and S6K2 are frequently coexpressed, and associated with a poor prognosis and endocrine resistance in breast cancer: a retrospective study including patients from the randomised Stockholm tamoxifen trials. *Breast Cancer Res.* 2013;15(5):R96