# Identification of a Methylation-Driven Gene Panel for Survival Prediction in Colon Cancer

**Yaojun Peng**
Chinese PLA General Hospital https://orcid.org/0000-0003-3652-2452

**Jing Zhao**
Department of Scientific Research Administration, The First Medical Centre, Chinese PLA General Hospital, Beijing, China

**Fan Yin**
Department of Oncology, The Second Medical Centre & National Clinical Research Center of Geriatric Disease, Chinese PLA General Hospital, Beijing, China

**Gaowa Sharen**
Department of Pathology, The First Affiliated Hospital of Inner Mongolia Medical University, Hohhot City, Inner Mongolia, China

**Qiyan Wu**
Department of Oncology, The First Medical Centre, Chinese PLA General Hospital, Beijing, China

**Xiaoxuan Sun**
National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital & Department of Oncology Surgery, Tianjin Cancer Hospital

**Juan Yang**
Department of Cardiothoracic Surgery, Tianjin 4th Center Hospital, Tianjin, China

**Huan Wang** ( wanghuanhuadian@163.com )
Department of Scientific Research Administration, The First Medical Centre, Chinese PLA General Hospital, Beijing, China https://orcid.org/0000-0002-1732-9203

---

**Research**

**Keywords:** Epigenetics, DNA methylation, Colon cancer, Prognosis, Integrative analyses, TCGA

**Posted Date:** June 30th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-37406/v1



---

# Abstract

**Background:** Prediction and improvement of prognosis is important for effective clinical management of colon cancer patients. Accumulation of a variety of genetic as well as epigenetic changes in colon epithelial cells has been identified as one of the fundamental processes that drive the initiation and progression of colon cancer. This study aimed to explore functional genes regulated by DNA methylation and the potential of these DNA methylation changes to become biomarkers predictive of colon cancer prognosis.

**Methods:** Methylation-driven genes (MDGs) were explored by applying an integrative analysis tool (MethylMix) to The Cancer Genome Atlas (TCGA) colon cancer project. TCGA colon cancer patients with available survival information (n=281) were randomly divided into training dataset (50%) for model construction and testing dataset (50%) for model validation. The prognostic MDG panel was identified in the training dataset by combining the Cox regression model with the least absolute shrinkage and selection operator regularization, a widely used approach to penalize the effect of multicollineatity. GSEA was employed to determine functional pathways associated with the prognostic 6-MDG panel. CD40 expression and methylation in colon cancer samples were also examined in datasets (expression profile [GSE8671] and methylation profile [GSE42752]) from Gene Expression Omnibus. Experimental confirmation of DNA methylation in colon cancer cell lines was performed using methylation specific PCR and bisulfite sequencing.

**Results:** We identified and internal validated a prognostic methylation-driven gene panel consisting of six gene members (TMEM88, HOXB2, FGD1, TOGARAM1, ARHGDIB and CD40). High risk phenotype classified by the 6-MDGs panel was associated with cancer-related biological processes, including invasion and metastasis, angiogenesis and tumor immune microenvironment, among others. The prognostic value of the 6-MDGs panel was independent of TNM stage, and its combination with TNM stage and age could help improve survival prediction of colon cancer patients. Additionally, we validated that the expression of CD40 was regulated by promoter region methylation in colon cancer samples and cell lines.

**Conclusions:** The proposed 6-MDGs panel represents a promising signature for estimating overall survival in patients with colon cancer.


# Background

Colorectal cancer (CRC) ranks third in cancer incidence and second in cancer-related mortality worldwide[1], with heterogeneous outcomes and distinct underlying pathobiologic and molecular features. Generalized screening of risky population with precursor initiating adenomas at age 50 years or older is an effective and durable strategy to find earlier staged cancers, reducing incidence and mortality of CRC[2-4]. Surgical resection of the primary cancer and/or limited metastasis is the only approach for attempted cure, with additional chemoradiation to improve outcome in some patients[4, 5]. However,

relapse or metachronous metastase occurs in a subset of these patients, leading to high mortality[6]. Therefore, robust diagnostic, prognostic and predictive biomarkers are clearly and urgently needed.

Currently, TNM staging is the only well-recognized stratification system used in clinical practice to guide therapy decision and predict CRC patients' prognosis[7, 8]. However, the fact that the survival time in patients with the same TNM stage of CRC is often heterogeneous highlights the need for more accurate strategies[9]. Genetic changes, such as gene mutations have long been known to contribute to cancer formation and used to predicts CRC patients' outcome[10, 11]. Recently, there is consensus that epigenetic alterations, including aberrant DNA methylation, abnormal histone modifications, and altered expression of non-coding RNAs occur early and manifest more frequently than genetic changes in CRC[12]. In addition, advances in genomic technologies and bioinformatics have led to the identification of specific epigenetic alterations as potential clinical biomarkers for CRC patients[12, 13]. For example, with the availability of genomic platforms capable of broadly surveying gene expression and DNA methylation, as evidenced by The Cancer Genome Atlas (TCGA) project, we can now identify genomic subtypes of CRCs[14, 15], and the CpG island methylator phenotype (CIMP) has undoubtedly been one of the most promising epigenetic biomarkers for prognosticating CRC patients[12, 16].

By applying an integrative analysis tool (MethylMix) to colon cancer samples from TCGA project, we sought to explore functional genes regulated by DNA methylation and the potential of these DNA methylation changes to become biomarkers predictive of colon cancer prognosis. We identified a prognostic methylation-driven gene (MDG) panel consisting of six gene members (TMEM88, HOXB2, FGD1, TOGARAM1, ARHGDIB and CD40). High risk phenotype classified by the 6-MDGs panel was associated with cancer-related biological processes, including invasion and metastasis, angiogenesis and tumor immune microenvironment, among others. We also confirmed expression and methylation of CD40, a member of the 6-MDG panel, in colon cancer samples and cell lines.

## Methods

# Data acquisition and preprocessing

Level 3 DNA methylation data of colon adenocarcinoma (COAD) samples measured by the Illumina Human Methylation 450 Beadchip (450K array) were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/) using the TCGA-Assembler[17]. These data are preprocessed via TCGA pipelines and presented in the form of b value, a ratio between methylated probe intensities and total probe intensities and probe-level data are condensed to a summary beta value by calculating the average methylation value for all CpG sites associated with a gene[18]. Totally, 353 samples of DNA methylation, including 315 COAD samples and 38 tumor adjacent samples were obtained. Methylation data were normalized using limma R package. Level 3 RNA-seq data and clinical information were also retrieved from the TCGA data portal. Among 521 cases of transcriptome profiles, 41 cases were obtained from tumor adjacent tissues, while the remaining 480 cases were COAD tissues. The transcriptome data were normalized and log2 transformed with the functions of DEGList and calcNormFactors in edgeR

package[19]. The clinical data were preprocessed by exclusion of samples without survival status and patients whose survival time was less than 30 days were also removed[20]. Two additional datasets of colon cancer (expression profile [GSE8671] and methylation profile [GSE42752]) downloaded from Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) were used to examine the expression and methylation of CD40, respectively. GSE8671 dataset contains transcriptional data of 32 COAD patients with adjacent normal mucosa from the same individuals evaluated by Affymetrix Human Genome U133 Plus 2.0 Array[21]. GSE42752 dataset includes genome-wide DNA methylation profile obtained from 22 COAD samples with corresponding adjacent normal colon mucosa and 20 cancer-unrelated healthy colon mucosa using 450K array[22]. Above data are available with no restrictions for research, and this study was performed in accordance with the guidelines of TCGA and GEO.

# Identification of MDGs

To identify MDGs, the MethylMix R package was applied to perform an analysis integrating gene expression and DNA methylation data. In the MethylMix algorithm, the methylation state of a gene is established by a b mixture model and hypo- or hyper-methylated genes are determined by comparing their differential methylation state in cancer versus normal tissues (false discovery rate [FDR] < 0.05)[23, 24]. MDGs should also fulfill the criteria of having a significant predictive effect on gene expression, implying that their methylation is inversely associated with transcription (Pearson coefficient < -0.3, P < 0.05) and thus functionally relevant[23, 24].

# Construction of the prognostic model for survival prediction

Survival analysis was performed on 281 COAD patients for whom both methylation and survival information (overall survival > 30 days) were available. We first randomly selected 50% of COAD patients as the training set and the remaining 50% of COAD patients as the testing set. Data matrixes were generated by combining methylation levels of the identified MDGs with matched follow-up data of COAD patients in the training set or the testing set. Then, univariate Cox regression analysis was performed to screen MDGs significantly associated with overall survival (P < 0.05) based on their methylation b value in the training set. Least absolute shrinkage and selection operator (LASSO) estimation, a well suited approach when there are a large number of correlated covariates in the patient cohort for model construction[25], was then performed to penalize the effect of multicollinearity using the glmnet R package[26]. MDGs survived from the LASSO estimation were subsequently subjected to multivariate Cox regression to construct a best fitting prognostic model with the Akaike information criterion (AIC) indicating model fitness[27]. Survival R package was used to execute steps in the univariate and multivariate Cox regression.

# Risk score calculation

The risk score was calculated by a linear combination of the methylation b value of the selected MDGs weighted by their estimated regression coefficient in the multivariable Cox regression analysis as discussed previously[28]. COAD patients were classified into high or low risk group using the median risk score of the training set as the cutoff value.

# Gene set enrichment analysis (GSEA)

GSEA[29] was employed to determine whether the members of a given gene set were generally associated with the risk score derived from the prognostic 6-MDG panel. In the whole process, the risk score (high or low) was designated as the phenotype and the analysis was conducted on the matched gene expression profile. Random sample permutations and the significant threshold were set at 1000 times and FDR < 0.01, respectively. GSEA was performed by the JAVA program (http://software.broadinstitute.org/gsea/index.jsp) using MSigDB C2 CP: KEGG gene set collection. The enriched KEGG pathways were ranked by normalized enrichment score (NES), and if a gene set have a positive NES, high expression level of the majority of its members is positively related to high risk score phenotype.

# Experimental validation in colon cancer cell lines

A panel of six colon cancer cell lines (RKO, SW480, SW620, HCT116, DLD1 and LoVo) were included in the present study. All these cell lines were preserved in our institute (The First Medical Centre, Chinese PLA General Hospital, Beijing, China) and were cultured in RPMI 1640 supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin.

Semi-quantitative RT-PCR to evaluate mRNA expression of CD40 in colon cancer cell lines with or without 5-aza-2’-deoxycytidine (5-Aza, Sigma) treatment (2 mM for 96h) were carried out as previously described[30]. Genomic DNA was prepared by the proteinase K method. Bisulfite treatment, methylation specific PCR (MSP) and bisulfite sequencing (BSSQ) were performed as previously described[31]. Genomic sequences around the transcriptional start site (TSS) were used as the template for CpG island prediction and design of MSP and BSSQ primers using Methyl Primer Express software v1.0 (Thermo Fisher Scientific). The primers for RT-PCR, MSP and BSSQ were listed in Table S1.

Total protein of CD40 in these colon cancer cell lines was measured by western blot, which was performed as previously described[30]. b-actin was used as a loading control. The antibodies used in western blot were purchased from the Proteintech company (Wuhan, China). We also examined membrane expression of CD40 using flow cytometry analysis. Cells were harvested using trypsin and washed with phosphate-buffered solution before incubation with or without the presence of phycoerythrin (PE)-tagged mouse monoclonal antibody to human CD40 (Sino Biological) at 4 °C for 30min. Then the cells were washed twice to remove unbound antibody before they were measured on a FACSCalibur flow cytometer (BD BioSciences).

# Statistical analysis

The Mann-Whitney test and Wilcoxon matched-pairs signed rank test were used to analyze the differences of DNA methylation, gene expression and risk score in non-paired and paired samples, respectively. The relativity between risk score and clinicopathological characteristics was analyzed using the Chi-square test or Fisher's exact test. Survival difference between the high-risk and low-risk group was evaluated by the Kaplan-Meier analysis, and log-rank test was used as a statistical method. Multivariate Cox regression and data stratification analysis were performed to test whether the risk score derived from the prognostic MDGs panel was independent of COAD patients' clinicopathological features. Receiver operating characteristic (ROC) curve was employed and the area under ROC curve (AUC) was calculated to compare the sensitivity and specificity of survival prediction based on age, TNM stage, the risk score derived from the prognostic 6-MDG panel and their combination. Statistical tests were conducted by GraphPad Prism8 (GraphPad Software) or R 3.6.0 using the corresponding R package mentioned above.

## Results

# Screening MDGs in COAD

We first prepared relevant expression and methylation data for the same patients and three data matrixes were acquired: a gene expression profile of 308 tumor tissues and two methylation profiles of 38 adjacent and 308 tumor tissues, respectively. These profiles were used as input data for MethylMix R package where methylation differential analysis and correlation analysis between DNA methylation and gene expression were conducted. According to the screen criteria, a total of 299 methylation-driven genes were identified (Table S2). The methylation profile of the most significant 30 hypo- and hyper-methylated MDGs (ranked by b value difference between tumor and adjacent tissues) was shown in Figure 1A. The correlations between DNA methylation and gene expression and methylation mixture models of the top 3 MDGs were shown in Figure 1B & 1C, respectively.

# Identification of a prognostic 6-MDG panel in the training set

A total of 281 COAD patients with both methylation and adequate follow up (survival time > 30 days) data were included in the survival analysis after preprocessing of the methylation and clinical data. The clinical information of these 281 COAD patients were summarized in Table S3. They were randomly split into the training set (n = 141) and the testing set (n = 140). To identify certain prognostic MDGs, univariate Cox regression analysis was performed in the training set and 12 prognosis related MDGs ($P < 0.05$; Table S4) were chosen for subsequent LASSO estimation. Ten MDGs survived the LASSO regularization (Figure 2A) after penalization of the multicollinearity effect and were further subjected to multivariable Cox regression analysis to construct a best fitting prognostic model. AIC was used to indicate model fitness. Finally, a prognostic DNA methylation gene panel consisting of six MDGs

(TMEM88, HOXB2, FGD1, TOGARAM1, ARHGDIB and CD40) was identified. The detailed information of the six MDGs was summarized in Table 1. The methylation profile, correlations between gene expression and DNA methylation and the methylation mixture models of the six MDGs were shown in Figure S1. The prognostic 6-MDG panel included one gene member (ARHGDIB) with statistically non-significant P value (P = 0.071; Table 1), but the 6-MDG panel acquired the lowest AIC representing the best model fitness and the overall effect was significant (AIC = 202.86, global P [Log-rank] < 0.001).

Next, a risk score model for overall survival prediction was created based on the methylation b values of these six MDGs, as follows: risk score = (-6.150* methylation b value of TMEM88) + (-3.593* methylation b value of HOXB2) + (-7.287* methylation b value of FGD1) + (-7.861* methylation b value of TOGARAM1) + (-3.622* methylation b value of ARHGDIB) + (-4.288* methylation b value of CD40). Then we calculated the risk score for each COAD patient, and classified them into high or low risk subgroup using the median risk score of the patients in the training set as the cutoff value.

Kaplan-Meier survival curve analysis of the training set showed that COAD patients in the high-risk group had significantly shorter median OS than those in the low-risk group (Log-rank P < 0.001; Figure 2B). We also profiled the distribution of risk score, survival status and methylation b value in the training set (Figure 2C-E). The risk scores of the patients in the training set ranged from -17.883 to -9.677 with the median risk score of -13.807 (Figure 2C). Moreover, there were more patients alive in the low-risk group than those in the high-risk group ($c^2$ = 13.45, P = 0.0002; Figure 2D). Interestingly, methylation levels of all the six MDGs were higher in low-risk patients than those in the high-risk patients (Figure 2E), indicating that hypermethylation of the 6-MDG panel is a favorable prognostic factor of COAD patients.

# The 6-MDG panel is predictive of survival in the testing and entire set

To further test the significance of the prognostic 6-MDG panel in COAD patients, the testing and entire set were used as validation groups. Using the same cutoff value of risk score obtained from the training set, COAD patients in the testing set were divided into high-risk group (n = 75) and low-risk group (n = 65). The result of Kaplan-Meier analysis demonstrated that COAD patients in the high-risk group showed worse overall survival than those in the low-risk group (Log-rank P = 0.0137; Figure 3A), and there were more patients alive in the low-risk group than those in the high-risk group ($c^2$ = 4.514, P = 0.0336; Figure 3B). We also performed the same analysis on the entire set (training set plus testing set, n = 281) and the results were consistent with those in the training and testing set (Figure 3C & 3D). Above results suggested that the 6-MDG panel can predict survival in both training and entire set.

## The prognostic value of the 6-MDG panel is independent of TNM stage

TNM classification is widely used, clinically useful and is highly associated with 5-year overall survival in colon cancer[32]. Thus, we set to clarify whether the prognostic value of the 6-MDG panel is independent of TNM stage. For this, we performed multivariable Cox regression and stratification analysis in the entire set. The multivariable Cox regression analysis was conducted on 271 patients, with age, gender, TNM stage and risk score as covariates. Ten cases of patients were excluded because of missing information on TNM stage. The results showed that age, TNM stage and the risk score remained independent prognostic factors in the multivariable Cox regression analysis (Figure 4A). Data stratification analysis was then performed where these patients were stratified into four subgroups (Stage I, II, III, and IV). The results of stratification analysis showed that the prognostic 6-MDG panel could identify patients with different overall survival in TNM stage II (Log-rank P = 0.0450) and IV (Log-rank P = 0.0160) subgroups (Figure 4B), while was insufficient to clarify the patients in TNM stage I (Log-rank P = 0.0750) and TNM stage III (Log-rank P = 0.0975) with significantly disparate survival (Figure 4B). This might be attributed to the small sample size or some truncated data. Thus, we combined low TNM stage (Stage I plus II) and high TNM stage (Stage III plus IV), and the risk score could significantly identify patients with different prognoses in these two subgroups (Log-rank P = 0.0083 and 0.0006, respectively; Figure 4C). Above results suggested that prognostic value of the 6-MDG panel is independent of TNM stage.

Moreover, we performed ROC analysis to compare the sensitivity and specificity of overall survival prediction between the prognostic factors including age, TNM stage, the risk score derived from the 6-MDG panel and combination of these three factors. As shown in Figure 4D, there was no significant difference when the AUCs of the three prognostic factors (age, TNM stage and the risk score) alone were pairwise compared (all $P > 0.05$). However, when these three prognostic were combined, the AUC was significantly greater than that of each prognostic factor alone (all $P < 0.05$). These results indicated that the combination of the three prognostic factors (age, TNM stage and the risk score) may help improve survival prediction in patients with COAD.

# Assessment of biological pathways associated with the 6-MDG panel

We performed GSEA to identify relevant pathways the 6-MDG might be involved in using the risk score for phenotype classification. Gene sets significantly enriched (FDR < 0.01) in the high-risk phenotype were shown in Figure 5A. The high risk score was positively associated with up-regulation of several cancer-related pathways, among others such as invasion and metastasis, angiogenesis and tumor immune microenvironment. For instance, vascular endothelial growth factor (VEGF), a key regulator of the growth and maintenance of blood vessels, can directly modulate the vascular wall by loosening cell-cell contacts and increasing the leakiness of blood vessels which favors tumor cell dissemination[33].

Next, we analyzed the relativity between the clinicopathological features and the risk score derived from the 6-MDG panel in COAD patients (Table 2). Consistent with the pathway analysis, the results showed that COAD patients in high-risk group were more likely to have remote metastasis ($c^2 = 6.465$, $P = 0.011$;

Table 2 & Figure 5B). We also evaluated the risk score as a continuous variable and patients with metastasis tended to have higher risk score than those without metastasis (P = 0.0036; Figure 5C). Collectively, above results suggested that the 6-MDG panel is associated with cancer-related signaling pathways and acts as an indicator of tumor metastasis.

# CD40 is universally hypermethylated in colon cancer tissues

CD40 is a member of the tumor necrosis factor family and a new immunemodulating target with great potential in cancer treatment[34]. Regulation of CD40 expression by DNA methylation has not been reported in current literatures, and thus deserves further investigation. We first inquired the expression of CD40 in colon cancer patients from the TCGA and GSE8671 dataset. The transcriptional expression of CD40 was significantly downregulated in colon cancer tissues compared with adjacent colon mucosa in both datasets (Figure 6A). Next, we analyzed the overall methylation level of CD40 in TCGA and GSE42725 dataset. The results showed that CD40 was hypermethylated in colon cancer tissues compared with adjacent or/and healthy colon mucosa in these two datasets (Figure 6B). We also observed a negative correlation between the mRNA expression and the overall DNA methylation level in COAD patients in TCGA (Pearson r = -0.511, P < 0.001; Figure S1B).

Additionally, we analyzed the CpG site-specific methylation status of all the 15 CpG sites of CD40 assessed by 450K array. The CpG sites located in or near the CpG island (Island, N shore and S shore) covering TSS of CD40 (12 CpG sites) were significantly hypermethylated in colon cancer tissues compared with the adjacent mucosa (Figure 6C), and their methylation levels were negatively correlated with CD40 expression, except for cg24575067 (Figure 6D). Interestingly, we observed a similar CpG site-specific methylation pattern of CD40 in the GSE42725 dataset (Figure 6E). Above results suggested that CD40 is universally hypermethylated in colon cancer tissues which may contribute to its transcriptional silence.

# The expression of CD40 is regulated by promoter methylation in colon cancer cell lines

To better understand the regulation of CD40 expression in colon cancer, the levels of CD40 expression were detected in a panel of colon cancer cell lines. CD40 mRNA expression was silenced in 3 of 6 colon cancer cell lines (Figure 7A). We confirmed the expression of CD40 using western blot and flow cytometry analysis on the total and membrane protein level in these six cell lines (Figure 7B & 7C). Next, MSP and BSSQ were employed to interrogate the methylation status of CD40 promoter region in these cell lines. The CpG island situated in CD40 gene promoter region and the designed MSP and BSSQ primers were shown in Figure 7D. MSP analysis revealed CD40 promoter methylation in the three cell lines (SW480, SW620 and DLD1) with silenced CD40 expression (Figure 7E). BSSQ analysis of 19 CpG sites around the TSS showed dense methylation in the CD40 silenced cell lines examined (SW480 and DLD1), but not in

the CD40 expressing HCT116 cells (Figure 7F). To test whether promoter methylation directly contributes to transcriptional silencing of CD40, these six colon cancer cell lines were treated with 5-Aza, a demethylation reagent. Restoration of CD40 expression was induced by 5-Aza in the three colon cancer cell lines with silenced CD40 expression (Figure 7G). These results indicated that CD40 is silenced in colon cancer cell lines by promoter region hypermethylation.

# Discussion

Aberrant epigenetic changes are crucial for carcinogenesis and subsequent tumor progression[35]. Of the various epigenetic modifications, DNA methylation acts as the key element and is classically responsible for transcriptional silence via hypermethylation of CpG islands located in promoter regions of tumor suppressor genes[36]. Additionally, DNA hypomethylation has been implicated in the regulation of genome rearrangement and chromosomal instability which may also contribute to carcinogenesis[36]. A plethora of gene-specific studies have demonstrated that hyper or hypomethylation of a gene can be utilized as epigenetic biomarker to predict behavior and prognosis of CRC[37]. There is also evidence of an association between aberrant methylation of multiple genes and increased CRC aggressiveness[38]. For instance, Weisenberger and colleagues later introduced the prevailing method used to identify CIMP in CRC, which is based on the methylation status of five genes, CACNA1G, IGF2, NEUROG1, RUNX3, and SOCS126[36]. CIMP-positive tumors exhibit unique clinicopathological and molecular features, correlating with an overall unfavorable prognosis[39].

The advance and prevalence of high through-put DNA methylation arrays have confirmed prior identified epigenetics changes, and on the other hand have uncovered a plenty of new alterations, creating an opportunity to discover novel cancer-related epigenetic biomarkers. By applying an integrative analysis tool to TCGA project, we sought to explore key genes regulated by DNA methylation and the potential of them to become prognostic biomarkers of colon cancer. A model-based algorithm (MethylMix) was employed to identify MDGs, based on which we developed a prognostic MDG panel consisting of six gene members (TMEM88, HOXB2, FGD1, TOGARAM1, ARHGDIB and CD40) in the training set (50% of TCGA cohort). The 6-MDG panel exhibited favorable performance on survival prediction, which was validated in the testing (the remaining 50% of TCGA cohort) and the entire set. Multivariate Cox regression and data stratification analysis demonstrated that the prognostic value of the risk score derived from the 6-MDG panel is independent of TNM stage. Furthermore, it was fascinating to find that the combination of age, TNM stage and the 6-MDG panel, three independently prognostic factors revealed by the multivariate Cox regression analysis might help improve prognosis prediction in the ROC curve analysis.

These six prognostic MDGs are differential methylated between tumor and adjacent tissues and their levels of DNA methylation and mRNA expression are inversely correlated. Such differentiation and relativity signify their potential roles in colon cancer. Our pathway analysis by GSEA provided evidence that the six MDGs are involved in cancer-related biological processes, including invasion and metastasis, angiogenesis and tumor immune microenvironment, among others. Up-regulation of HOXB2 was found

to be an adverse prognostic indicator for stage I lung adenocarcinomas, promoting invasion by transcriptional regulation of metastasis-related genes[40, 41]. In our study, expression of HOXB2 is negatively correlated with DNA methylation in colon cancer and hypermethylation of HOXB2 is associated with prolonged overall survival. However, Marsit et al. revealed a distinct role of HOXB2 in bladder cancer that increased promoter methylation of HOXB2 is significantly and independently associated with higher degree of cancer aggressiveness[42]. Thus, further studies are greatly needed to clarify the functional role of HOXB2 in cancer. ARHGDIB has been identified as a regulator of tumor metastasis but its role in cancer remains controversial[43]. ARHGDIB can function as a positive (in ovary[44], breast[45], colorectal[43] and gastric cancer[46]) and negative (in Hodgkin's lymphoma[47], bladder[48, 49] and lung cancer[50]) regulator of cancer progression. In our study, hypermethylation of ARHGDIB is a favorable prognostic factor, in agreement with previous findings in colon cancer. TMEM88 is a transmembrane protein and functions as an inhibitor of Wnt signaling[51]. TMEM88 promoter hypomethylation is associated with platinum resistance in ovarian cancer[52]. Our study demonstrated that TMEM88 is hypomethylated in the high-risk group with shorter overall survival in colon cancer. So we hypothesis that TMEM88 may modulate the prognosis of colon cancer via altering sensitivity of cancer cell to chemodrug mediated by promoter methylation, and further investigations are needed to confirm that. Ayala et al. revealed a central role for FGD1 in regulating focal degradation of the extracellular matrix at invadopodia[53]. They also demonstrated that FGD1 is highly expressed in prostate and breast cancer, which might lead to aberrant growth, invasiveness, and/or metastatic potential[53]. TOGARAM1 encodes a TOG domain array-containing protein that regulates cilia microtubule structure[54]. Regulation of TOGARAM1 expression by DNA methylation and its role in cancer has not been reported. CD40 belongs to the family of TNF receptors and is crucial in mediating a variety of immune and inflammatory responses[55]. CD40 ligation provides essential activation signals for immune cells[55], while CD40 possesses controversial functions in promoting or inhibiting tumorigenesis and progression via regulation of TNFα-induced apoptosis[56], angiogenesis[57], tumor cell migration and invasion[58] and chemoresistance[59]. Agonist CD40 antibodies have been developed and tested in clinical trials where impressive results have been noted, especially in pancreatic cancers[60]. We confirmed that the expression of CD40 is regulated by promoter region hypermethylation in colon cancer tissues and cell lines, which may bring new sight into the combination of epigenetic therapy and CD40-stimulating immunotherapy. Further investigations are warranted to clarify the underlying mechanisms that potentiate above methylation-driven genes as DNA methylation biomarkers for colon cancer.

Several limitations should be acknowledged for our study. First, no external validation was performed. We attempted to search for colon cancer cohorts with both methylation and follow-up data in multiple cancer databases, including GEO and the International Cancer Genome Consortium (ICGC) project, among others, but no available dataset was found. However, considering the sufficient number of patients included in the process of model construction and internal validation, the identified prognostic signature is unlikely to be a random noise of the methylome. Second, limited experimental information regarding to the regulatory mechanisms of all the six prognostic MDGs in the methylation signature was presented. Third, the specific functional role of these prognostic MDGs in colon cancer was left unexplored.

## Conclusion

In summary, we identified a MDG-related signature which acts as an independent prognostic factor in colon cancer and its combination with clinical characters including age and TNM stage could help improve prognosis prediction. We also confirmed that CD40, a member of the prognostic 6-MDG panel, is regulated by DNA methylation in colon cancer samples and cell lines. More efforts are necessary to have a complete picture of the regulatory mechanisms and functional roles of all the six MDGs in colon cancer. Also clinical investigations in additional colon cancer patient cohorts are warranted to validate our findings and to elaborate its potential utility.

## Abbreviations

CRC: Colorectal cancer; TCGA: The Cancer Genome Atlas; CIMP: CpG island methylator phenotype; MDG: Methylation-driven gene; COAD: Colon adenocarcinoma; 450K array: Illumina Human Methylation 450 Beadchip; GEO: Gene Expression Omnibus; FDR: False discovery rate; LASSO: Least absolute shrinkage and selection operator; AIC: Akaike information criterion; NES: Normalized enrichment score; MSP: Methylation specific PCR; BSSQ: Bisulfite sequencing; TSS: Transcriptional start site; 5-Aza: 5-aza-2'-deoxycytidine; ROC: Receiver operating characteristic; AUC: Area under receiver operating characteristic; VEGF: vascular endothelial growth factor; ICGC International Cancer Genome Consortium.

## Declarations

# Ethics approval and consent to participate

All procedures performed in this study, involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments. Consent for participation for all patients was obtained through The Cancer Genome Atlas project or the corresponding original work where the datasets were generated.

# Consent for publication

Not applicable.

# Availability of data and materials

Clinical information, high-throughput sequencing-counts and DNA methylation data were retrieved from the TCGA data portal (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/geo/), which are publicly available databases.

# Competing interests

The authors declare that there is no conflict of interests.

# Funding

This work was supported by National Natural Science Foundation of China (81972902), Translational Medicine Program of Chinese PLA General Hospital (2017TM-022) and Youth Talents Promotion Project (17-JCJQ-QT-030).

# Authors' contributions

All authors contributed to the experimental design and the analysis of data in this study. YP, JZ, FY, GS and QW downloaded, organized and analyzed the data. YP, XS and JY performed validation experiments in colon cell lines and wrote the draft. HW supervised this study and revised the manuscript. All authors have read and commented on the manuscript and approved the final version.

# Acknowledgements

None.

# References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA Cancer J Clin* 2018, **68:**394-424.
2. Nishihara R, Ogino S, Chan AT: **Colorectal-cancer incidence and mortality after screening.** *N Engl J Med* 2013, **369:**2355.
3. Shaukat A, Mongin SJ, Geisser MS, Lederle FA, Bond JH, Mandel JS, Church TR: **Long-term mortality after screening for colorectal cancer.** *N Engl J Med* 2013, **369:**1106-1114.
4. Carethers JM, Jung BH: **Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer.** *Gastroenterology* 2015, **149:**1177-1190 e1173.
5. Boland CR, Sinicrope FA, Brenner DE, Carethers JM: **Colorectal cancer prevention and treatment.** *Gastroenterology* 2000, **118:**S115-128.
6. Chang W, Gao X, Han Y, Du Y, Liu Q, Wang L, Tan X, Zhang Q, Liu Y, Zhu Y, et al: **Gene expression profiling-derived immunohistochemistry signature with high prognostic value in colorectal carcinoma.** *Gut* 2014, **63:**1457-1467.
7. Haggar FA, Boushey RP: **Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors.** *Clin Colon Rectal Surg* 2009, **22:**191-197.

8. Xiong Y, Wang R, Peng L, You W, Wei J, Zhang S, Wu X, Guo J, Xu J, Lv Z, Fu Z: **An integrated lncRNA, microRNA and mRNA signature to improve prognosis prediction of colorectal cancer.** *Oncotarget* 2017, **8:**85463-85478.
9. Zinicola R, Pedrazzi G, Haboubi N, Nicholls RJ: **The degree of extramural spread of T3 rectal cancer: a plea to the UICC and AJCC.** *Colorectal Dis* 2017, **19:**310.
10. Sinicrope FA, Okamoto K, Kasi PM, Kawakami H: **Molecular Biomarkers in the Personalized Treatment of Colorectal Cancer.** *Clin Gastroenterol Hepatol* 2016, **14:**651-658.
11. Van Schaeybroeck S, Allen WL, Turkington RC, Johnston PG: **Implementing prognostic and predictive biomarkers in CRC clinical trials.** *Nat Rev Clin Oncol* 2011, **8:**222-232.
12. Okugawa Y, Grady WM, Goel A: **Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers.** *Gastroenterology* 2015, **149:**1204-1225 e1212.
13. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL: **Principles and methods of integrative genomic analyses in cancer.** *Nat Rev Cancer* 2014, **14:**299-313.
14. Cancer Genome Atlas N: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487:**330-337.
15. Lee MS, Menter DG, Kopetz S: **Right Versus Left Colon Cancer Biology: Integrating the Consensus Molecular Subtypes.** *J Natl Compr Canc Netw* 2017, **15:**411-419.
16. Ogino S, Cantor M, Kawasaki T, Brahmandam M, Kirkner GJ, Weisenberger DJ, Campan M, Laird PW, Loda M, Fuchs CS: **CpG island methylator phenotype (CIMP) of colorectal cancer is best characterised by quantitative DNA methylation analysis and prospective cohort studies.** *Gut* 2006, **55:**1000-1006.
17. Zhu Y, Qiu P, Ji Y: **TCGA-assembler: open-source software for retrieving and processing TCGA data.** *Nat Methods* 2014, **11:**599-600.
18. Sanford T, Meng MV, Railkar R, Agarwal PK, Porten SP: **Integrative analysis of the epigenetic basis of muscle-invasive urothelial carcinoma.** *Clin Epigenetics* 2018, **10:**19.
19. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
20. Kruppa J, Jung K: **Automated multigroup outlier identification in molecular high-throughput data using bagplots and gemplots.** *BMC Bioinformatics* 2017, **18:**232.
21. Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E, Maake C, Rehrauer H, Laczko E, Kurowski MA, Bujnicki JM, Menigatti M, et al: **Transcriptome profile of human colorectal adenomas.** *Mol Cancer Res* 2007, **5:**1263-1275.
22. Naumov VA, Generozov EV, Zaharjevskaya NB, Matushkina DS, Larin AK, Chernyshov SV, Alekseev MV, Shelygin YA, Govorun VM: **Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips.** *Epigenetics* 2013, **8:**921-934.
23. Gevaert O: **MethylMix: an R package for identifying DNA methylation-driven genes.** *Bioinformatics* 2015, **31:**1839-1841.

24. Gevaert O, Tibshirani R, Plevritis SK: **Pancancer analysis of DNA methylation-driven genes using MethylMix.** *Genome Biol* 2015, **16:**17.
25. Tibshirani R: **Regression shrinkage and selection via the lasso: a retrospective.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011, **73:**273-282.
26. Engebretsen S, Bohlin J: **Statistical predictions with glmnet.** *Clin Epigenetics* 2019, **11:**123.
27. Harrell FE, Jr., Lee KL, Mark DB: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.** *Stat Med* 1996, **15:**361-387.
28. Pan Y, Song Y, Cheng L, Xu H, Liu J: **Analysis of methylation-driven genes for predicting the prognosis of patients with head and neck squamous cell carcinoma.** *J Cell Biochem* 2019, **120:**19482-19495.
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
30. Guo Y, Peng Y, Gao D, Zhang M, Yang W, Linghu E, Herman JG, Fuks F, Dong G, Guo M: **Silencing HOXD10 by promoter region hypermethylation activates ERK signaling in hepatocellular carcinoma.** *Clin Epigenetics* 2017, **9:**116.
31. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB: **Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands.** *Proc Natl Acad Sci U S A* 1996, **93:**9821-9826.
32. Dienstmann R, Mason MJ, Sinicrope FA, Phipps AI, Tejpar S, Nesbakken A, Danielsen SA, Sveen A, Buchanan DD, Clendenning M, et al: **Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study.** *Ann Oncol* 2017, **28:**1023-1031.
33. Saharinen P, Eklund L, Pulkki K, Bono P, Alitalo K: **VEGF and angiopoietin signaling in tumor angiogenesis and metastasis.** *Trends Mol Med* 2011, **17:**347-362.
34. Piechutta M, Berghoff AS: **New emerging targets in cancer immunotherapy: the role of Cluster of Differentiation 40 (CD40/TNFR5).** *ESMO Open* 2019, **4:**e000510.
35. Jones PA, Baylin SB: **The epigenomics of cancer.** *Cell* 2007, **128:**683-692.
36. Egger G, Liang G, Aparicio A, Jones PA: **Epigenetics in human disease and prospects for epigenetic therapy.** *Nature* 2004, **429:**457-463.
37. Coppede F, Lopomo A, Spisni R, Migliore L: **Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer.** *World J Gastroenterol* 2014, **20:**943-956.
38. Sakai E, Nakajima A, Kaneda A: **Accumulation of aberrant DNA methylation during colorectal cancer development.** *World J Gastroenterol* 2014, **20:**978-987.
39. Juo YY, Johnston FM, Zhang DY, Juo HH, Wang H, Pappou EP, Yu T, Easwaran H, Baylin S, van Engeland M, Ahuja N: **Prognostic value of CpG island methylator phenotype among colorectal cancer patients: a systematic review and meta-analysis.** *Ann Oncol* 2014, **25:**2314-2327.
40. Inamura K, Togashi Y, Ninomiya H, Shimoji T, Noda T, Ishikawa Y: **HOXB2, an adverse prognostic indicator for stage I lung adenocarcinomas, promotes invasion by transcriptional regulation of**

**metastasis-related genes in HOP-62 non-small cell lung cancer cells.** *Anticancer Res* 2008, **28:**2121-2127.

41. Inamura K, Togashi Y, Okui M, Ninomiya H, Hiramatsu M, Satoh Y, Okumura S, Nakagawa K, Shimoji T, Noda T, Ishikawa Y: **HOXB2 as a novel prognostic indicator for stage I lung adenocarcinomas.** *J Thorac Oncol* 2007, **2:**802-807.
42. Marsit CJ, Houseman EA, Christensen BC, Gagne L, Wrensch MR, Nelson HH, Wiemels J, Zheng S, Wiencke JK, Andrew AS, et al: **Identification of methylated genes associated with aggressive bladder cancer.** *PLoS One* 2010, **5:**e12334.
43. Li X, Wang J, Zhang X, Zeng Y, Liang L, Ding Y: **Overexpression of RhoGDI2 correlates with tumor progression and poor prognosis in colorectal carcinoma.** *Ann Surg Oncol* 2012, **19:**145-153.
44. Tapper J, Kettunen E, El-Rifai W, Seppala M, Andersson LC, Knuutila S: **Changes in gene expression during progression of ovarian carcinoma.** *Cancer Genet Cytogenet* 2001, **128:**1-6.
45. Zhang Y, Zhang B: **D4-GDI, a Rho GTPase regulator, promotes breast cancer cell invasiveness.** *Cancer Res* 2006, **66:**5592-5598.
46. Cho HJ, Baek KE, Park SM, Kim IK, Choi YL, Cho HJ, Nam IK, Hwang EM, Park JY, Han JY, et al: **RhoGDI2 expression is associated with tumor growth and malignant progression of gastric cancer.** *Clin Cancer Res* 2009, **15:**2612-2619.
47. Ma L, Xu G, Sotnikova A, Szczepanowski M, Giefing M, Krause K, Krams M, Siebert R, Jin J, Klapper W: **Loss of expression of LyGDI (ARHGDIB), a rho GDP-dissociation inhibitor, in Hodgkin lymphoma.** *Br J Haematol* 2007, **139:**217-223.
48. Seraj MJ, Harding MA, Gildea JJ, Welch DR, Theodorescu D: **The relationship of BRMS1 and RhoGDI2 gene expression to metastatic potential in lineage related human bladder cancer cell lines.** *Clin Exp Metastasis* 2000, **18:**519-525.
49. Theodorescu D, Sapinoso LM, Conaway MR, Oxford G, Hampton GM, Frierson HF, Jr.: **Reduced expression of metastasis suppressor RhoGDI2 is associated with decreased survival for patients with bladder cancer.** *Clin Cancer Res* 2004, **10:**3800-3806.
50. Said N, Sanchez-Carbayo M, Smith SC, Theodorescu D: **RhoGDI2 suppresses lung metastasis in mice by reducing tumor versican expression and macrophage infiltration.** *J Clin Invest* 2012, **122:**1503-1518.
51. Ge YX, Wang CH, Hu FY, Pan LX, Min J, Niu KY, Zhang L, Li J, Xu T: **New advances of TMEM88 in cancer initiation and progression, with special emphasis on Wnt signaling pathway.** *J Cell Physiol* 2018, **233:**79-87.
52. de Leon M, Cardenas H, Vieth E, Emerson R, Segar M, Liu Y, Nephew K, Matei D: **Transmembrane protein 88 (TMEM88) promoter hypomethylation is associated with platinum resistance in ovarian cancer.** *Gynecol Oncol* 2016, **142:**539-547.
53. Ayala I, Giacchetti G, Caldieri G, Attanasio F, Mariggio S, Tete S, Polishchuk R, Castronovo V, Buccione R: **Faciogenital dysplasia protein Fgd1 regulates invadopodia biogenesis and extracellular matrix degradation and is up-regulated in prostate and breast cancer.** *Cancer Res* 2009, **69:**747-752.

54. Das A, Dickinson DJ, Wood CC, Goldstein B, Slep KC: **Crescerin uses a TOG domain array to regulate microtubules in the primary cilium.** *Mol Biol Cell* 2015, **26:**4248-4264.
55. van Kooten C, Banchereau J: **CD40-CD40 ligand.** *J Leukoc Biol* 2000, **67:**2-17.
56. Tewari R, Choudhury SR, Mehta VS, Sen E: **TNFalpha regulates the localization of CD40 in lipid rafts of glioma cells.** *Mol Biol Rep* 2012, **39:**8695-8699.
57. Xie F, Shi Q, Wang Q, Ge Y, Chen Y, Zuo J, Gu Y, Deng H, Mao H, Hu Z, et al: **CD40 is a regulator for vascular endothelial growth factor in the tumor microenvironment of glioma.** *J Neuroimmunol* 2010, **222:**62-69.
58. Zhou Y, Zhou SX, Gao L, Li XA: **Regulation of CD40 signaling in colon cancer cells and its implications in clinical tissues.** *Cancer Immunol Immunother* 2016, **65:**919-929.
59. Yamaguchi H, Tanaka F, Sadanaga N, Ohta M, Inoue H, Mori M: **Stimulation of CD40 inhibits Fas- or chemotherapy-mediated apoptosis and increases cell motility in human gastric carcinoma cells.** *Int J Oncol* 2003, **23:**1697-1702.
60. Vonderheide RH: **The Immune Revolution: A Case for Priming, Not Checkpoint.** *Cancer Cell* 2018, **33:**563-569.
61. Ma R, Feng N, Yu X, Lin H, Zhang X, Shi O, Zhang H, Zhang S, Li L, Zheng M, et al: **Promoter methylation of Wnt/beta-Catenin signal inhibitor TMEM88 is associated with unfavorable prognosis of non-small cell lung cancer.** *Cancer Biol Med* 2017, **14:**377-386.
62. Nagata H, Kozaki KI, Muramatsu T, Hiramoto H, Tanimoto K, Fujiwara N, Imoto S, Ichikawa D, Otsuji E, Miyano S, et al: **Genome-wide screening of DNA methylation associated with lymph node metastasis in esophageal squamous cell carcinoma.** *Oncotarget* 2017, **8:**37740-37750.
63. Xavier FC, Destro MF, Duarte CM, Nunes FD: **Epigenetic repression of HOXB cluster in oral cancer cell lines.** *Arch Oral Biol* 2014, **59:**783-789.
64. Cai C, Xie X, Zhou J, Fang X, Wang F, Wang M: **Identification of TAF1, SAT1, and ARHGEF9 as DNA methylation biomarkers for hepatocellular carcinoma.** *J Cell Physiol* 2020, **235:**611-618.
65. Wang L, Shi J, Huang Y, Liu S, Zhang J, Ding H, Yang J, Chen Z: **A six-gene prognostic model predicts overall survival in bladder cancer patients.** *Cancer Cell Int* 2019, **19:**229.
66. Huang T, Yang J, Cai YD: **Novel candidate key drivers in the integrative network of genes, microRNAs, methylations, and copy number variations in squamous cell lung carcinoma.** *Biomed Res Int* 2015, **2015:**358125.
67. Zeller C, Dai W, Steele NL, Siddiq A, Walley AJ, Wilhelm-Benartzi CS, Rizzo S, van der Zee A, Plumb JA, Brown R: **Candidate DNA methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling.** *Oncogene* 2012, **31:**4567-4576.

# Tables

Table 1. Six individual genes of the methylation-driven gene panel associated with overall survival of colon cancer patients.

| Gene symbol | Description | Chr | Coefficient | P value | Associated with DNA methylation in cancer |
|---|---|---|---|---|---|
| TMEM88 | Transmembrane protein 88 | 17p13.1 | -6.150 | 0.018 | OC[52] and NSCLC[61] |
| HOXB2 | Homeobox B2 | 17q21.32 | -3.593 | 0.001 | ESSC[62], OSCC[63] and BC[42] |
| FGD1 | Rho GEF and PH domain containing 1 | Xp11.22 | -7.287 | 0.003 | HCC[64] |
| TOGARAM1 | TOG array regulator of axonemal microtubules 1 | 14q21.2 | -7.861 | 0.042 | NR |
| ARHGDIB | Rho GDP dissociation inhibitor beta | 12p12.3 | -3.622 | 0.071 | BC[65], LUSC[66] and OC[67] |
| CD40 | Cluster of differentiation 40 | 20q13.12 | -4.288 | 0.004 | NR |

OC, Ovarian cancer; NSCLC, non-small cell lung cancer; ESSC, esophageal squamous cell carcinoma; OSCC, oral squamous cell carcinoma; BC, bladder cancer; HCC, hepatocellular carcinoma; OC, ovarian cancer; LUSC, lung squamous cell carcinoma; NR: Not reported.

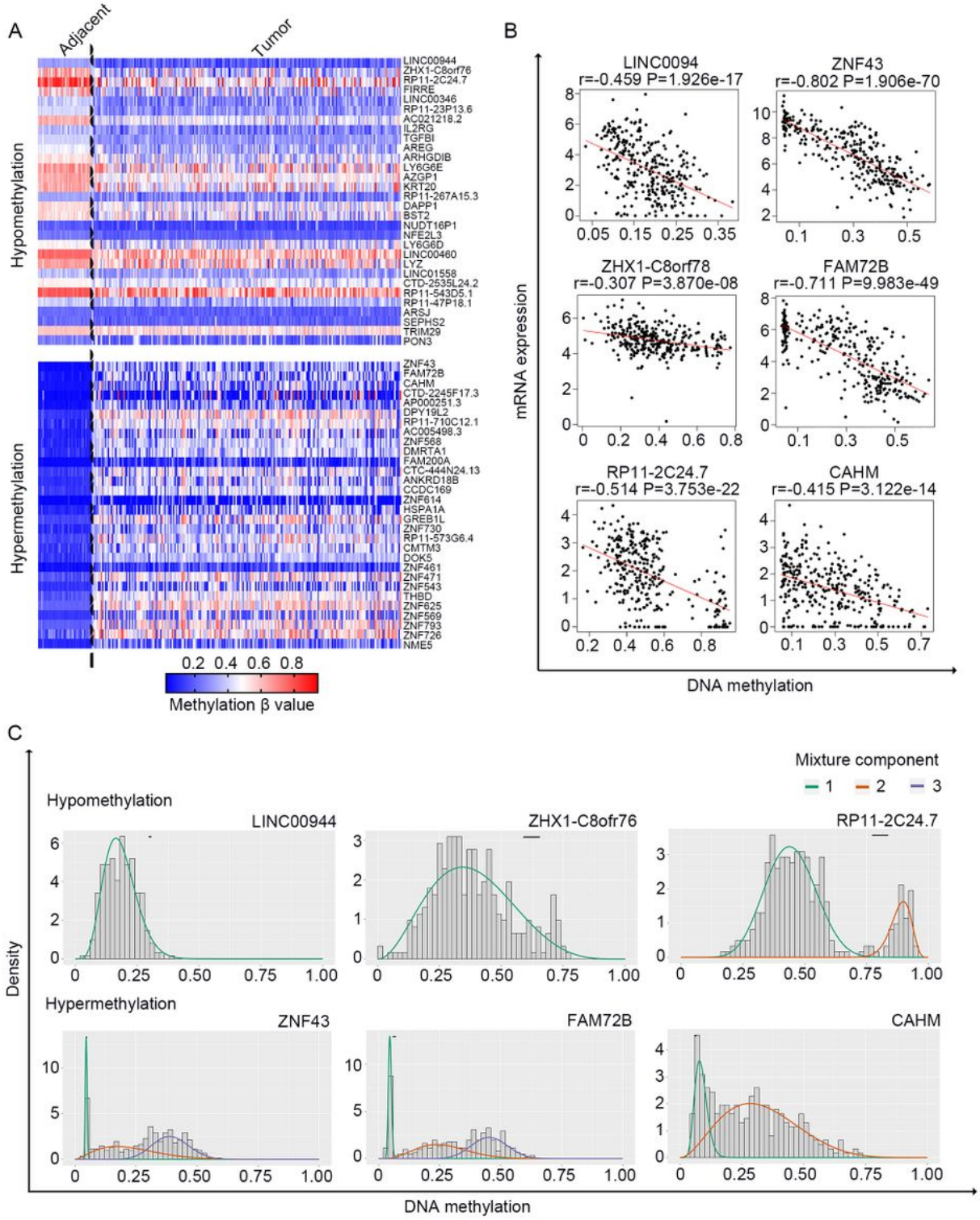
Table 2. Correlations between clinicopathological features and risk score derived from the six methylation-driven gene panel.

| Variable | N | High risk | Low risk | P value |
|---|---|---|---|---|
| Age (years) | 281 | | | 0.535 |
| >=60 | 185 | 93 | 92 | |
| <60 | 96 | 52 | 44 | |
| Gender | 281 | | | 0.002* |
| Male | 153 | 92 | 61 | |
| Female | 128 | 53 | 75 | |
| History of colon polyps | 213 | | | 0.304 |
| Yes | 50 | 21 | 29 | |
| No | 163 | 82 | 81 | |
| Pretreatment CEA level (ng/ml) | 184 | | | 0.067 |
| >=5.0 | 61 | 38 | 23 | |
| <5.0 | 123 | 59 | 64 | |
| T stage | 281 | | | 0.236 |
| T3+T4 | 231 | 123 | 108 | |
| T1+T2 | 50 | 22 | 28 | |
| N stage | 281 | | | 0.886 |
| N1+N2 | 119 | 62 | 57 | |
| N0 | 162 | 83 | 79 | |
| M stage | 232 | | | 0.011* |
| M1 | 40 | 28 | 12 | |
| M0 | 192 | 92 | 100 | |
| TNM stage | 271 | | | 0.570 |
| □+□ | 122 | 64 | 58 | |
| □+□ | 149 | 73 | 76 | |
| Venous invasion | 243 | | | 0.267 |
| Yes | 58 | 34 | 24 | |

| | | | | |
|---|---|---|---|---|
| No | 185 | 93 | 92 | |
| Tumor location | 262 | | | 0.322 |
| Right colon | 164 | 80 | 84 | |
| Left colon | 98 | 54 | 44 | |

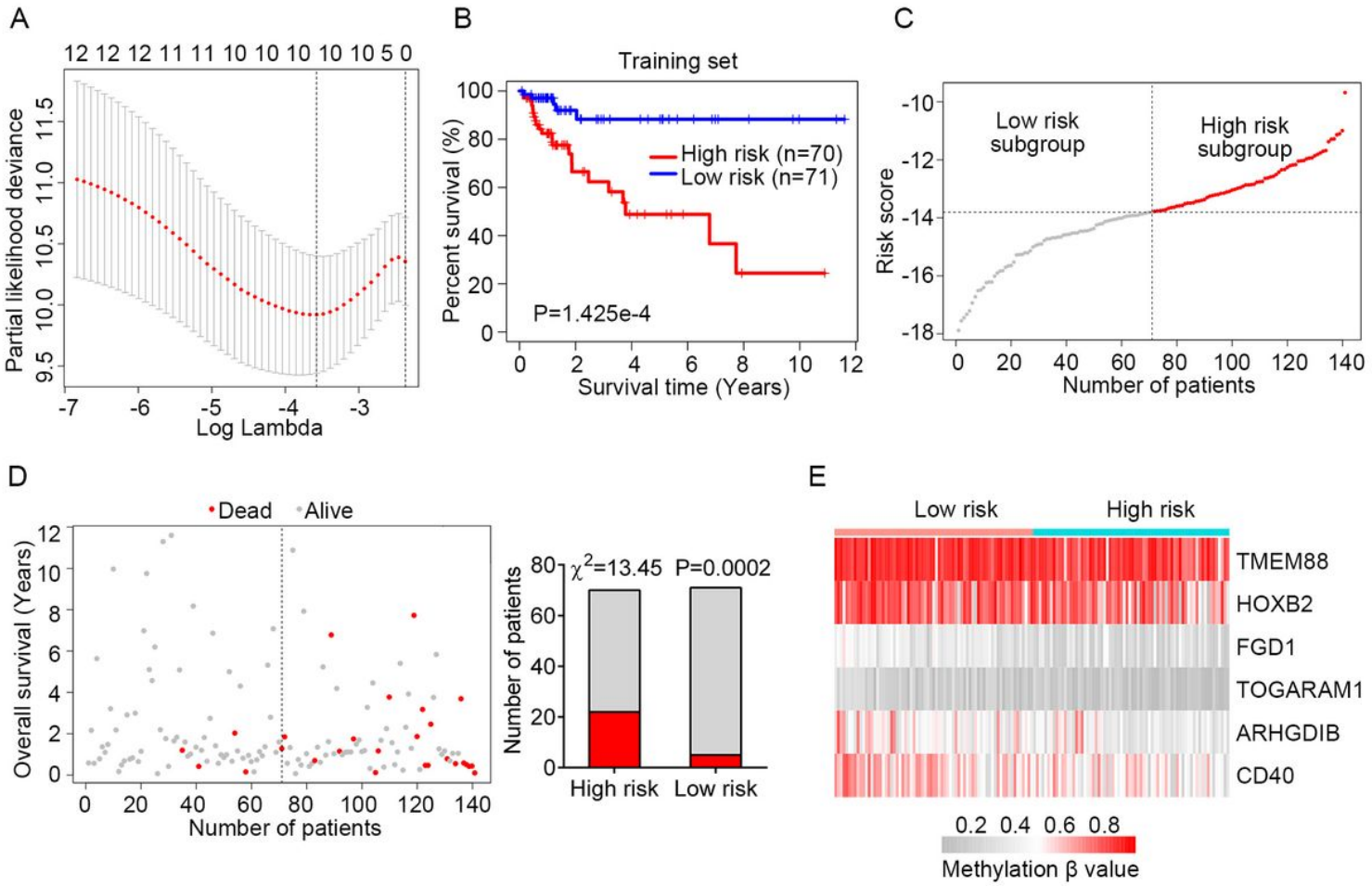
*P < 0.05; CEA, carcinoembryonic antigen.

# Figures



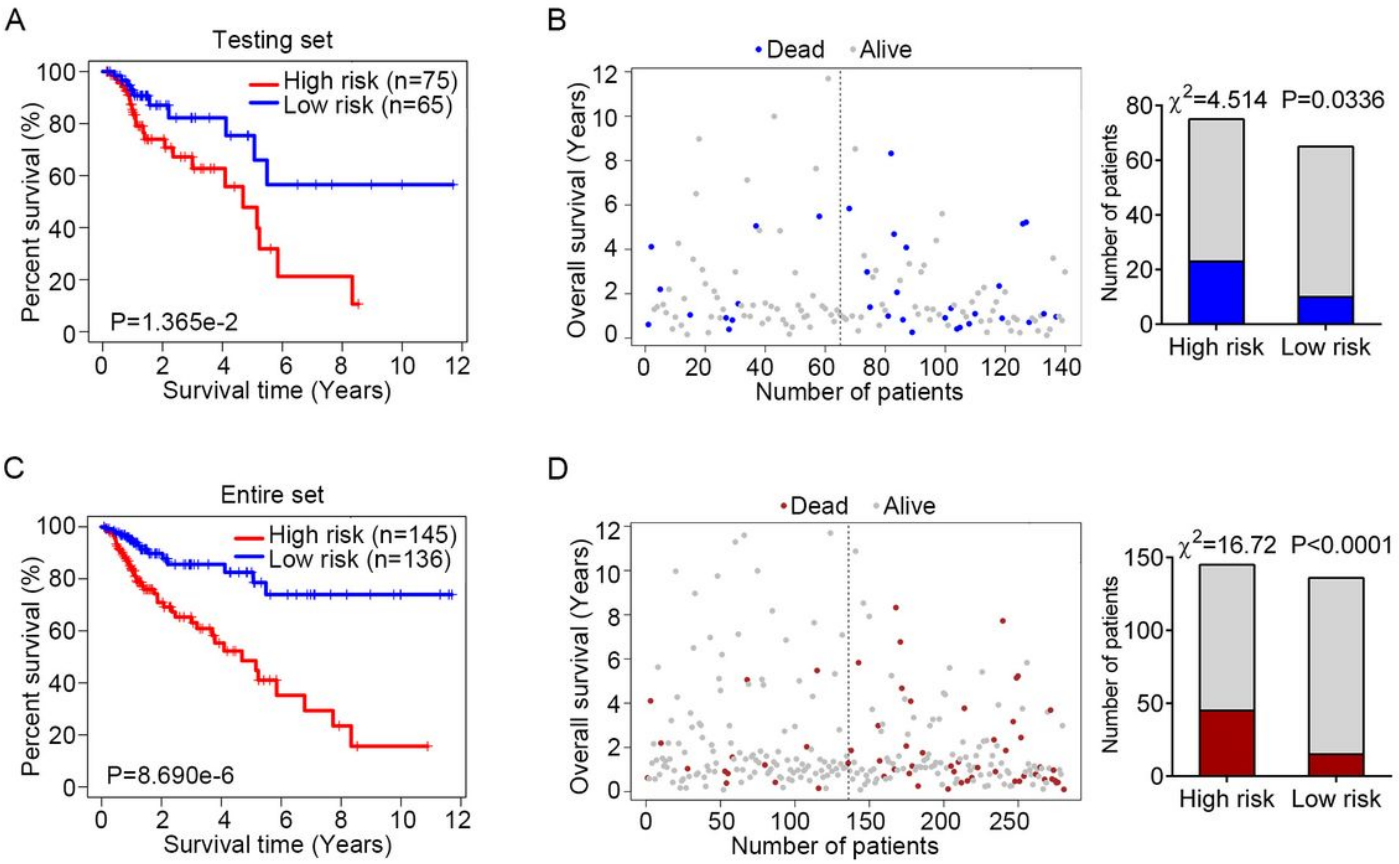

## Figure 1

Screening MDGs in colon cancer. (A) The methylation profile of the most significant 30 hypo- and hyper-methylated MDGs in adjacent and colon cancer tissues. (B) The association between gene expression and DNA methylation of the top three hypo- and hyper-methylated MDGs in colon cancer samples. (C) The mixture models of the top three hypo- and hyper-methylated MDGs in colon cancer samples. The horizontal black bar represents the distribution of methylation values in the normal samples.



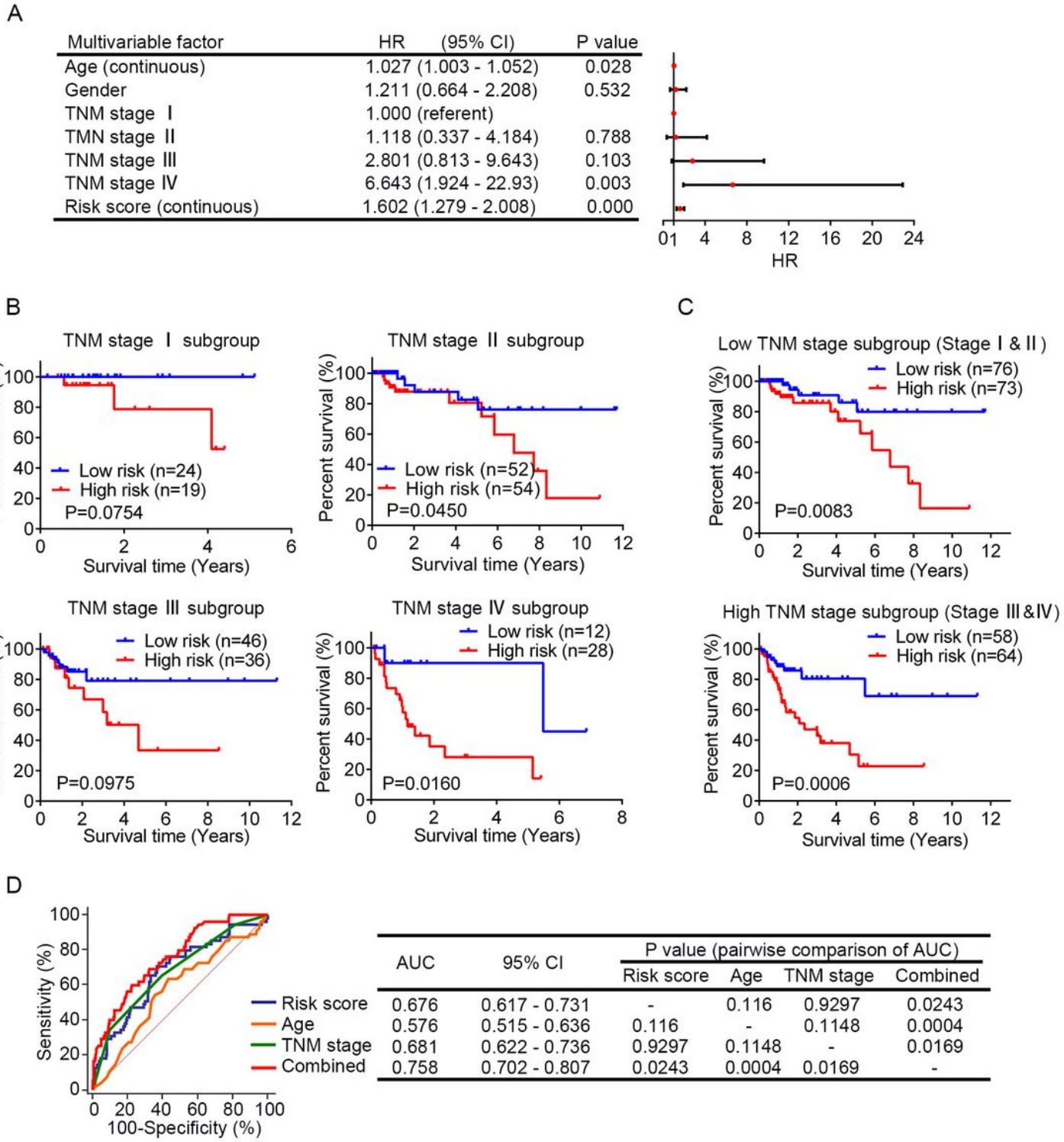

## Figure 2

Identification of a prognostic 6-MDG panel in the training set. (A) Ten-fold cross-validation for tuning parameter selection in the LASSO model. The dotted vertical line (left) is drawn at the optimal value by minimum criteria. (B) Kaplan-Meier estimate of the overall survival using the 6-MDGs panel in the training set. Colon cancer patients were divided into high-risk (n = 70) or low-risk (n = 71) subgroup based on the median value of risk score. The difference between the two curves was determined by the two-side log-rank test. (C) The distribution of risk score derived from the 6-MDGs in the training set. (D) The distribution of colon cancer patients' survival status in the training set. The difference between the high-risk and low-risk subgroup was determined by Chi-square test. (E) The methylation profile of the six MDGs in the training set.




## Figure 3

The 6-MDG panel is predictive of survival in the testing and entire set. (A) Kaplan-Meier estimate of the overall survival using the 6-MDGs panel in the testing set. Colon cancer patients were divided into high-risk (n = 75) or low-risk (n = 65) subgroup based on the median risk score of the training set. The difference between the two curves was determined by the two-side log-rank test. (B) The distribution of colon cancer patients' survival status in the testing set. The difference between the high-risk and low-risk subgroup was determined by Chi-square test. (C) Kaplan-Meier estimate of the overall survival using the 6-MDGs panel in the entire set. (D) The distribution of colon cancer patients' survival status in the entire set.

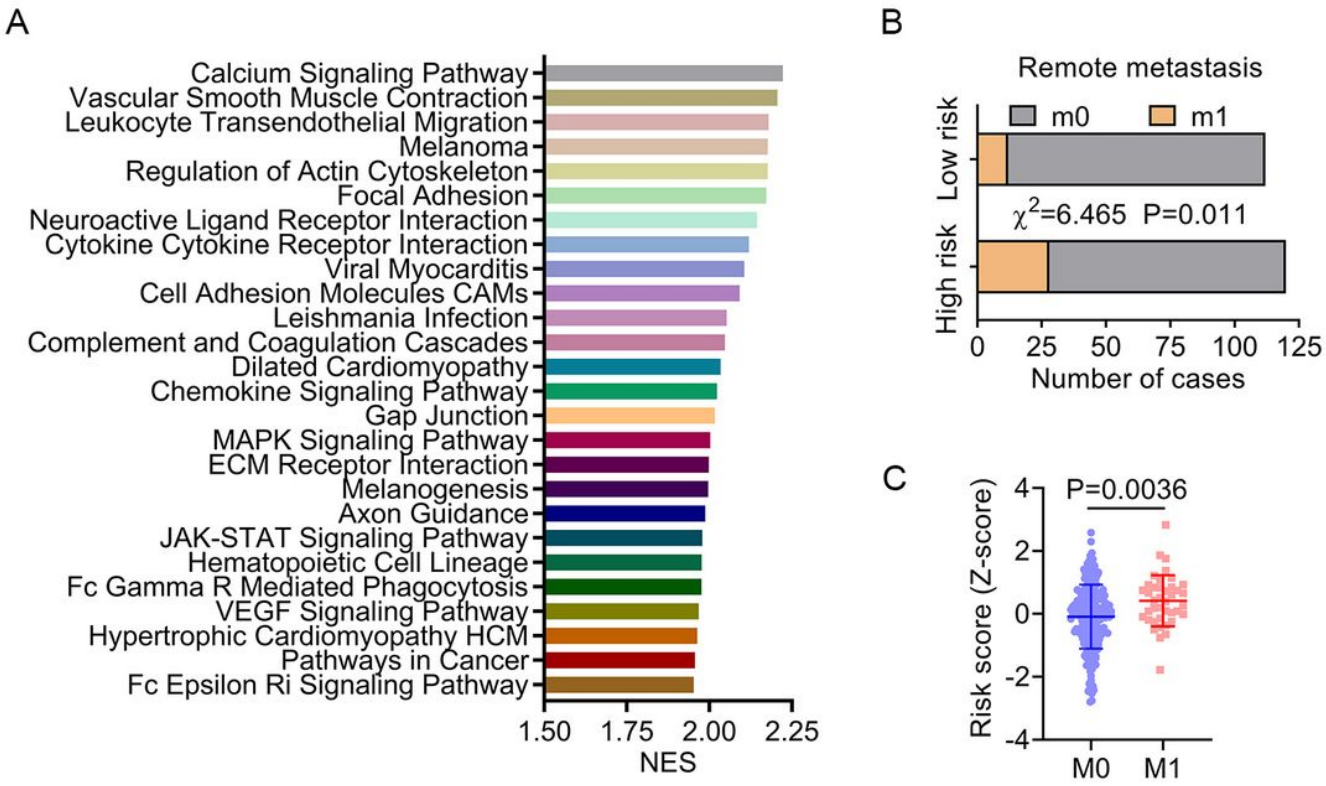



| AUC | 95% CI | P value (pairwise comparison of AUC) | | | |
|---|---|---|---|---|---|
| | | Risk score | Age | TNM stage | Combined |
| 0.676 | 0.617 - 0.731 | - | 0.116 | 0.9297 | 0.0243 |
| 0.576 | 0.515 - 0.636 | 0.116 | - | 0.1148 | 0.0004 |
| 0.681 | 0.622 - 0.736 | 0.9297 | 0.1148 | - | 0.0169 |
| 0.758 | 0.702 - 0.807 | 0.0243 | 0.0004 | 0.0169 | - |

## Figure 4

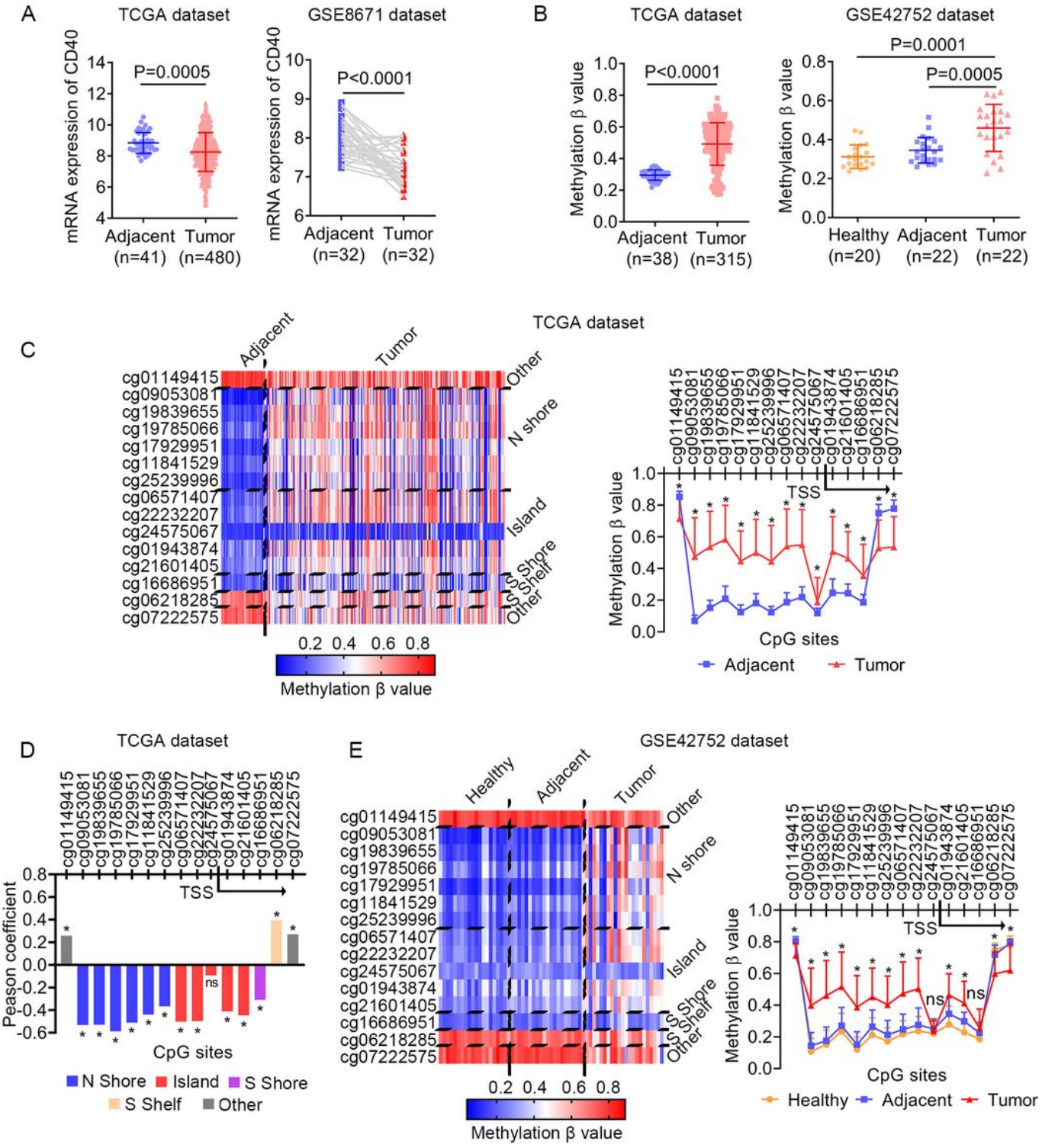
The Prognostic value of the 6-MDG panel is independent of TNM stage. (A) The multivariate Cox regression analysis performed on 271 colon cancer patients that contained age, gender, TNM stage and risk score as covariates. Risk score and age were evaluated as continuous variables, and gender and TNM stage were evaluated as category variables. Red solid dots represent the hazard ratio (HR) of death and open-ended horizontal lines represent the 95% confidence intervals (CIs). All P values were calculated

using Cox proportional hazards analysis. (B) Kaplan-Meier curves for colon cancer patients with TNM stage I (n = 43; upper left panel), TNM stage II (n = 106; upper right panel), TNM stage IV (n = 40; bottom left panel) and TNM stage III (n = 82; bottom left panel). The difference between the two curves was determined by the two-side log-rank test. (C) Kaplan-Meier curves for patients with low TNM stage (Stage I & II, n = 149; upper panel) and high TNM stage (Stage III & IV, n = 122; bottom panel). (D) ROC analysis of the sensitivity and specificity of overall survival prediction by age, TNM stage, the risk score derived from the 6-MDG panel and combination of above three factors. P values were obtained from the pairwise comparisons of the AUCs.




## Figure 5

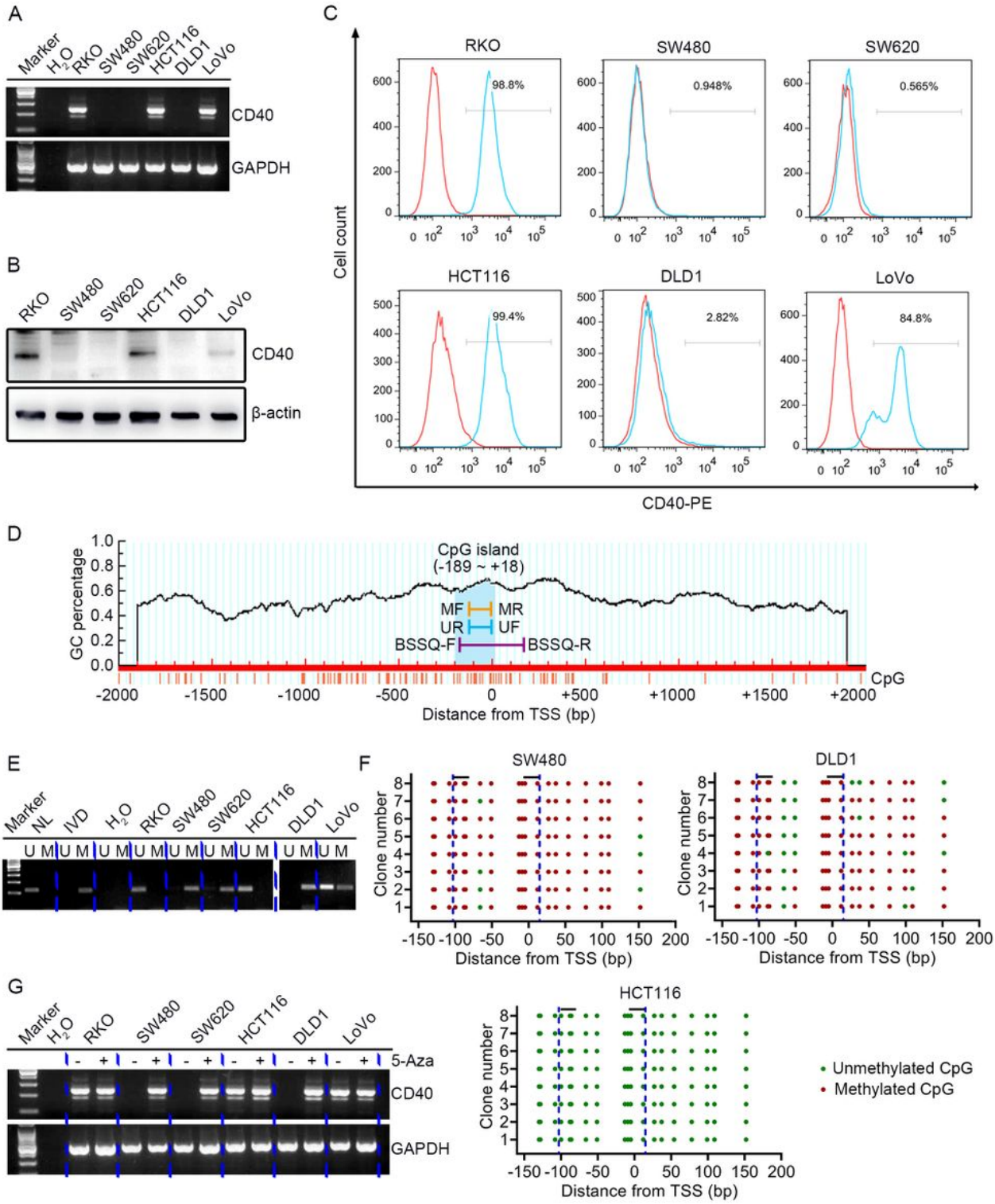
Assessment of relevant pathways and biological processes of the 6-MDG panel. (A) GSEA analysis showed significantly enriched KEGG pathways in colon cancer tissues with high risk phenotype (FDR < 0.01). (B) The relativity between remote metastasis and the risk score derived from the 6-MDG panel in colon cancer patients. The statistical difference was determined by Chi-square test. (C) Scatter plot of risk score of patients with or without metastasis. The Mann-Whitney test was used to determine the significance of the comparison.




## Figure 6

CD40 is universally hypermethylated in colon cancer tissues. (A) Scatter plots of CD40 mRNA expression between adjacent and colon cancer tissues from TCGA and GSE8671 dataset. The Mann-Whitney test was used to determine the significance of the comparison. (B) Scatter plots of CD40 DNA methylation ( value) between adjacent and colon cancer tissues from TCGA dataset and among healthy, adjacent and colon cancer tissues from GSE42752 dataset. The Mann-Whitney test and Wilcoxon matched-pairs

signed rank test were used to analyze the differences between non-paired and paired samples, respectively. (C) The methylation profile of all the CpGs (n=15) of CD40 in colon cancer samples from TCGA dataset. The differences of CpG sites' methylation levels between adjacent and tumor tissues were determined by the Mann-Whitney test. (D) Pearson correlations between CD40 mRNA expression and methylation levels of all the 15 CpG sites. (E) The methylation profile of all the CpGs (n=15) of CD40 in colon cancer samples from the GSE42752 dataset. The differences of CpG sites' methylation levels between adjacent and tumor tissues were determined by the Mann-Whitney test. TSS, transcriptional start site; *P < 0.05; ns, no significance.




## Figure 7

The expression of CD40 is regulated by promoter methylation in colon cancer cell lines. (A) mRNA, (B) total protein and (C) membrane expression of CD40 in six colon cancer cell lines (RKO, SW480, SW620, HCT116, DLD1 and LoVo). (D) Schematic diagram of CpG island in the promoter region of CD40. MF, methylation forward primer; MR, methylation reverse primer; UF, unmethylation forward primer; UR, unmethylation reverse primer; BSSQ-F, bisulfite sequencing forward primer; BSSQ-R, bisulfite sequencing

reverse primer. (E) Methylation status of CD40 detected by MSP in colon cancer cell lines. IVD, in vitro methylated DNA; NL, normal lymphocyte DNA; M, methylated alleles; U, unmethylated alleles. (F) BSSQ of CD40 performed in SW480, DLD1 and HCT116 cell lines. Red solid dots represent methylated CpG sites, and green solid dots denote unmethylated CpG sites. The horizontal black bar demarcates the primers of MSP, which are included in the region of BSSQ. (G) mRNA expression of CD40 with (+) or without (-) treatment of 5-Aza.

## Supplementary Files

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

- TableS1.docx
- TableS3.docx
- TableS2.docx
- TableS4.docx
- figs1300dpi.tif