A Prognostic Model Based on Nine DNA Methylation-Driven Genes Predicts Overall Survival for Colorectal Cancer

Zhongsheng Feng, Zhanju Liu, Kangsheng Peng* and Wei Wu*

Department of Gastroenterology, Shanghai Tenth People's Hospital, Tongji University, Shanghai, China

Background: Colorectal cancer (CRC) is the third most frequently diagnosed malignancy and the fourth leading cause of cancer-related death among common tumors in the world. We aimed to establish and validate a risk assessment model to predict overall survival (OS) for the CRC patients.

Methods: DNA methylation-driven genes were identified by integrating DNA methylation profile and transcriptome data from The Cancer Genome Atlas (TCGA) CRC cohort. Then, a risk score model was built based on LASSO, univariable Cox and multivariable Cox regression analysis. After analyzing the clinicopathological factors, a nomogram was constructed and assessed. Another cohort from GEO was used for external validation. Afterward, the molecular and immune characteristics in the two risk score groups were analyzed.

Results: In total, 705 methylation-driven genes were identified. Based on the LASSO and Cox regression analyses, nine genes, i.e., LINC01555, GSTM1, HSPA1A, VWDE, MAGEA12, ARHGAP, PTPRD, ABHD12B and TMEM88, were selected for the development of a risk score model. The Kaplan–Meier curve indicated that patients in the low-risk group had considerably better OS (P = 2e-08). The verification performed in subgroups demonstrated the validity of the model. Then, we established an OS-associated nomogram that included the risk score and significant clinicopathological factors. The concordance index of the nomogram was 0.81. A comprehensive molecular and immune characteristics analysis showed that the high-risk group was associated with tumor invasion, infiltration of immune cells executing pro-tumor suppression (such as myeloid-derived suppressor cells, regulatory T cells, immature dendritic cells) and higher expression of common inhibitory checkpoint molecules (ICPs).

Abbreviations: AUC, area under the curve; CRC, colorectal cancer; DEGs, differentially expressed genes; GEPIA, gene expression profiling interactive analysis database; GSEA, gene Set enrichment analysis; FDR, false discovery rate; ICPs, inhibitory checkpoint molecules; MDSCs, myeloid-derived suppressor cells; NES, normalized enrichment score; OS, overall survival; ROC, receiver operating characteristic curve; TIICs, tumor-infiltrating immune cells; TAMs, tumor-associated macrophages; WGCNA, weighted gene correlation network analysis.
INTRODUCTION

Colorectal cancer (CRC) is the third most frequently diagnosed malignancy and the fourth leading cause of cancer-related death among of those common tumors in the world (Arnold et al., 2017). The morbidity of CRC worldwide is expected to increase to more than two million new cases by 2035 (Dekker et al., 2019). Curative surgical resection and chemotherapy are still the main treatment (Dekker et al., 2019). Although great progress has been made in the diagnosis and treatment of CRC in recent years, the prognosis of colorectal cancer remains unsatisfactory. Currently, the prognosis assessment of colorectal cancer staging based on tumor-node-metastases (TNM) classification system is insufficient for prognostic estimation, which limited the clinical decision-making (Compton, 2007). The exploration of effective biomarkers for early assessment is an important preventive measure to improve the prognosis of CRC.

Pursuing predictors of colorectal cancer, a growing number of studies have identified valuable biomarkers, such as Integrin beta-4 (ITGB4) (Li et al., 2019), Placenta-specific protein 1 (PLAC1) (Ren et al., 2020), some miRNAs and IncRNAs (Hibner et al., 2018; Dastmalchi et al., 2020). However, due to lack of sufficient sensitivity and specificity, few biomarkers have been successfully applied in clinical practice.

DNA methylation represents an important epigenetic modification that regulates gene expression (Dor and Cedar, 2018). In humans, DNA methylation mainly occurs on CpG dinucleotides. CRC develops through cumulative genetic and epigenetic changes in the precursor lesions (such as adenomas and serrated lesions) (Jung et al., 2020). Currently, hypermethylation in promoter regions of some important tumor suppressor genes that cause gene expression inhibition has been identified in colorectal cancer cells (de Vallière et al., 2015), such as Cyclin-dependent kinase inhibitor 2A (CDKN2A) (Bihl et al., 2012), DNA mismatch repair protein Mlh1 (MLH1) (Cunningham et al., 1998), and Adenomatous polyposis coli protein (APC) (Liang et al., 2017). The aberrant methylation may be a key event in the progression of colorectal cancer. Therefore, deregulated DNA methylation and corresponding gene expression changes might be used as biomarkers to play a prospective role in the early diagnosis, prognosis and clinical decision-making for CRC, which is worthy of further study.

By integrating DNA methylation and mRNA expression profile data, we aimed to screen out CRC-associated DNA methylation-driven genes and evaluate the potential of these genes expression to predict CRC prognosis. We identified a prognosis-related gene panel made up of nine genes and developed a risk score model. Then we established a nomogram by combining the nine-gene signature and some significant clinicopathologic factors to predict overall survival (OS) in patients with CRC. The model was validated in one Gene Expression Omnibus (GEO) cohort. Furthermore, we characterized the molecular and immune profile of nine-gene signature. These results showed that the nine-gene risk model was a promising prognostic biomarker for CRC patients.

Conclusion: Our nine-gene associated risk assessment model is a promising signature to distinguish the prognosis for CRC patients. It is expected to serve as a predictive tool with high sensitivity and specificity for individualized prediction of OS in the patients with CRC.

Keywords: nomogram, risk score, colorectal cancer, DNA methylation, prognosis

MATERIALS AND METHODS

Data Acquisition and Preprocessing

The available RNA-seq transcriptome data, gene mutation information and clinicopathological information were directly downloaded from the TCGA database (https://portal.gdc.cancer.gov/). There were 476 samples with gene transcriptome data (41 normal and 435 tumor), 347 patients with mutation information, and 400 patients with available survival data. TCGA-Assembler 2 (Wei et al., 2018) was used to obtain Level 3 methylation data (37 normal and 279 tumor) from the TCGA Methylation 450 K Bead chip. Then the function CalculateSingleValueMethylationData of TCGA-Assembler two was used to calculate the average methylation level of each gene. The methylation levels of genes were scored using β values ranging from 0 to 1 (unmethylated to totally methylated). The external validation cohort was from a GEO dataset of gene expression arrays [GSE39582 (N = 550)] (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582). “Deseq2” package of R was used to normalize the raw RNA-seq data and identify differentially expressed genes (DEGs) between normal and cancer groups (Anders and Huber, 2010).

Identification of DNA Methylation-Driven Genes

Through the integration of gene expression and DNA methylation datasets, “MethylMix” package of R software was employed to recognize DNA methylation-driven genes (Gevaert, 2015). There are three steps. Firstly, a linear regression model was built to estimate the association between gene methylation and gene expression, and genes with a significant inverse relationship (p value < 0.01) were selected. Such genes were defined as transcriptionally predictive genes. Secondly, a beta mixture model was created to determine the methylation states of each gene. Thirdly, the Wilcoxon rank tests were computed to compare the methylation levels between each methylation state and normal tissue samples. Differential genes were collected. Lastly, genes that were both transcriptionally predictive and differential were selected as DNA methylation-driven genes.
Function Enrichment and Pathway Analysis
Metascape (Zhou et al., 2019) was used to perform GO Biological Processes enrichment and KEGG pathway analysis of the genes identified by MethylMix. Only terms with \( p < 0.01 \), a minimum overlap of 3 and an enrichment factor > 1.5 were considered significant.

Feature Selection and Building the Predictive Signature
Initially, the genes identified by MethylMix were applied to a univariable Cox regression and a LASSO Cox regression. In univariable Cox regression analysis, genes with \( p < 0.05 \) were selected. In LASSO regression, genes were screened 1,000 times, and if specific genes were detected more than 700 times, they were regarded as candidates. The intersection of univariable Cox regression and LASSO regression (nine genes) were brought into a multivariable Cox regression analysis. The linear combination of the regression coefficient derived from the multivariable Cox regression model (\( \beta \)) multiplied by its mRNA level generated a prognostic risk score with nine genes.

Development and Validation of the Risk Score Model
Using X-tile software (Camp et al., 2004) to determine the optimal cut-off value, we divided patients into low- and high-risk groups. Then the Kaplan–Meier curves and time-dependent receiver operating characteristic curve (tdROC) analysis were performed to evaluate the predictive accuracy of the nine-gene signature risk model for OS. The survivalROC function in “survivalROC” package of R was used to draw tdROC and the span argument was set at 0.0657. A subgroup analysis was performed by dividing the patients based on clinicopathological characteristics.

Development and Assessment of the Nomogram in the TCGA Dataset
The significance of the risk score model and other traditional clinicopathological characteristics to predict OS was evaluated by univariable Cox regression analysis. Then, we used multivariable Cox regression analysis and stepwise regression method to distinguish significant predictive factors, from which we built a nomogram predictive model. The concordance index (C-index) was computed to evaluate the accuracy of the nomogram. The prognostic risk value of each patient was calculated using the nomogram and tdROC curve analysis was used to further validate the predictive performance of the nomogram. The span argument of survivalROC function was set as 0.0616. The calibration of model was visualized in calibration plot and quantified by the slope of the calibration line.

Weighted Gene Correlation Network Analysis
All robustly expressed genes (nearly 18,000 genes) were used for WGCNA analysis. The analysis was performed as described previously (Langfelder and Horvath, 2008). Briefly, we first screened the best soft threshold by "WGCNA" package (Langfelder and Horvath, 2008) of R to ensure a scale-free network. In the co-expression network, genes with high absolute correlations were clustered into the same module. Furthermore, module eigengenes (MEs) were defined as the first principal component of each gene module and the expression of MEs was considered as a representative of all genes in a given module. The correlation between MEs and clinical trait was calculated to identify the clinical key module. GO and KEGG enrichment analysis of the key module resulted from WGCNA were performed by Metascape. ModuleMembership (MM) was defined as the Pearson's correlation between gene expression and ME. Gene significance (GS) was defined as the Pearson’s correlation between gene expression and certain clinical trait. The cut-off criteria was set as \([|MM|>0.8, |GS|>0.3]\) to identify hub genes with high connectivity in the clinical key module. Gene Expression Profiling Interactive Analysis (GEPIA) database (Tang et al., 2017) (http://gepia.cancer-pku.cn/) is an interactive website application that includes the transcriptome data in TCGA and GTEx projects and integrate them in a widely accepted process. We inputted the hub genes into the GEPIA and validated these hub genes.

Gene Set Enrichment Analysis
Differential expression analysis was first performed on all genes to analyze the samples with high and low nine-gene score using “DESeq2” package of R. Enrichment analysis to determine the signaling pathways in which the differentially expressed genes are involved was then carried out by using GSEA method based on the GO Biological Processes and Hallmark gene sets with the “clusterProfiler” (Yu et al., 2012) package of R. A false discovery rate (FDR) less than 0.25 and an absolute value of the normalized enrichment score (NES) greater than 1 were defined as the cutoff criteria.

Gene Mutation Analysis
In the gene mutation analysis, information on genetic alterations was directly obtained from the TCGA database portal, and the quantity and quality of gene mutations were analyzed by using the “Maftools” package (Mayakonda et al., 2018) of R.

Analysis of Tumor-Infiltrating Immune Cells Characteristics
The ssGSEA by “GSVA” package (Hänzelmann et al., 2013) of R was introduced to quantify the relative infiltration of 28 immune cell types in the tumor microenvironment. Feature gene sets for
each immune cell type were obtained from a recent publication (Charoentong et al., 2017). The gene sets include 782 genes for predicting the abundance of 28 TIICs in individual tissue samples. The following 28 types of immune cells include: activated B cells (Ba), activated CD4$^+$ T cells (CD4$^+$ Ta), activated CD8$^+$ T cells (CD8$^+$ Ta), activated dendritic cells (DCa), CD56bright natural killer cells (CD56$^+$ NK), CD56dim natural killer cells (CD56$^-$ NK), central memory CD4$^+$ T cells (CD4$^+$ Tcm), central memory CD8$^+$ T cells (CD8$^+$ Tcm), effector memory CD4$^+$ T cells (CD4$^+$ Tem), effector memory CD8$^+$ T cells (CD8$^+$ Tem), eosinophils, gamma delta T cells (γδT), immature B cells (Bim), immature dendritic cells (DCim), mast cells, myeloid-derived suppressor cells (MDSC), memory B cells (Bm), monocytes, natural killer cells (NK), natural killer T cells (NK T), neutrophils, plasmacytoid dendritic cells (DCp), macrophages, regulatory T cells (Tregs), follicular helper T cells (Tfh), type-1 T helper cells (Th1), type-17 T helper cells (Th17), and type-2 T helper cells (Th2). The relative abundance of each immune cell type was represented by an enrichment score in ssGSEA analysis. The ssGSEA score was normalized to unity distribution, for which zero is the minimal and one is the maximal score for each immune cell type.

Statistical Analysis
The bilateral log-rank test was used in the Kaplan–Meier survival examination to compare OS between different subgroups. Wilcoxon rank-sum test was used to evaluate the differences between two groups. All statistical analyses were conducted with R software (version 4.0.3). All statistical tests were two-sided, and $p$ values less than 0.05 were considered statistically significant.

RESULTS
TCGA Data Acquisition and Identification of DNA Methylation-Driven Genes
The study flowchart describing the process is shown in Figure 1. To identify DNA methylation-driven genes in CRC, we performed MethylMix analysis on 273 tumor samples and 37 normal tissue samples. A total of 705 genes were identified and a mixture model of each gene was constructed (Supplementary Figure S1). Their methylation levels were visualized by heat map (Figure 2A). Metascape analysis showed the first 20 clusters of enriched sets (Figure 2B). For biological process (BP), these genes showed enrichment in pattern specification process, lipid catabolic process, multicellular organismal homeostasis, cell
fate commitment and so on. The KEGG pathway data were enriched in peroxisome and PPAR signaling pathway.

Establishment of a Prognostic Risk Score Model

We included 400 patients with RNA-seq data and complete clinical information from the TCGA database for subsequent analysis. The patients’ characteristics are summarized in Table 1. Univariable Cox regression analysis was performed for 705 genes and 40 genes with \( p < 0.05 \) were selected. Besides, the same 705 genes were brought into LASSO regression analysis and 10 genes were selected (Figures 3A,B). The intersection of above two gene sets, including nine genes (LINC01555, GSTM1, HSPA1A, VWDE, MAGEA12, ARHGAP4, PTPRD, ABHD12B, TMEM88) were eventually screened out as prognosis-related genes. Supplementary Figure S2 showed the correlation between gene expression and corresponding methylation level of nine genes. Then, multivariable Cox regression analyses were performed, and a nine-gene model was constructed according to their expression levels and coefficients (Figure 3C). The
Development and Evaluation of a Nomogram for OS Prediction

Risk score, age, T stage, lymph node metastatic and distant metastatic were selected as significant predictive factors after univariable regression analysis (Table 2). Stepwise regression analysis showed that after removing the factor lymph node metastatic, AIC of multivariable Cox regression went down from 751.65 to 750.64, so it was further ruled out. Therefore, four independent risk factors (risk score, age, T stage and distant metastatic) were ultimately retained to build a visualized and applicable nomogram (Figures 5A,B). The C-index of the model is 0.81 (95% CI: 0.754–0.868). According to Nomogram, each variable was assigned a corresponding score, as a result we could calculate a total score for each patient. The AUC of nomogram was 0.815 (95% CI 0.712–0.917), 0.794 (95% CI 0.708–0.881) and 0.802 (95% CI 0.720–0.884) at 1-, 3- and 5-year, respectively (Figure 5C). The results revealed that the predicted survival possibility by the nomogram was close to the actual survival situation. Furthermore, our nomogram model performed better than the model only using significant clinicopathological factors (Supplementary Table S1).

External Validation of the Nomogram

The nomogram established above was further validated in the GEO dataset GSE39582 (Figures 6A–E). Totally 550 patients with complete basic clinical information were included for analysis (Table 1). The calibration curves for the predicted possibility of 1-, 3- and 5-years survival displayed obvious concordance between the predicted results and the actual observations in the GEO dataset (Figures 6B–D). The calibration slope was 1.56 (95% CI 1.31–1.82) at 1-year survival, 1.21 (95% CI 1.05–1.36) at 3-years survival and 1.00 (95% CI 0.74–1.26) at 5-years survival. Similar to the performance in TCGA cohort, the AUC of tdROC was 0.788 (95% CI 0.684–0.892) at 1-year survival, 0.743 (95% CI 0.679–0.807) at 3-years survival and 0.714 (95% CI 0.647–0.780) at 5-years survival (Figure 6E). The concordance index of the nomogram was 0.722 (95% CI 0.683–0.761). To sum up, the results showed that our model performed well in the validation cohort.

Molecular Characteristic of Different Risk Groups

We used two approaches to comprehensively explore molecular characteristic of different risk groups. Firstly, we performed WGCNA analysis in the TCGA cohort. In current study, the power of $\beta = 10$ (scale-free $R^2 = 0.74$) was selected as a soft threshold parameter to ensure a scale-free network (Figures 7A,B). The co-expression gene modules were then constructed and divided into 28 meaningful modules via the average linkage hierarchical clustering (Figure 7C). Blue module was found to have the highest association with 9-gene risk model (Cor = 0.28, $P = 1e-08$ for risk score; Cor = 0.2, $P = 6e-05$ for risk group) (Figure 7D). There were a total of 1,577 genes in the blue module. We then used Metascape to analyze these genes. Top 20 clusters of functional enriched sets were presented (Figure 8A). Based on

| Variables | Groups | Patients |
|-----------|--------|---------|
| | Training set ($n = 400$) | Validation set ($n = 550$) |
| Age [n (%)] | Median | 68 | 68 |
| | Range | 57.7–77.0 | 59.0–76.0 |
| | ≤65 years | 166 (41.5%) | 219 (39.8%) |
| | >65 years | 234 (58.5%) | 331 (60.2%) |
| Gender [n (%)] | Male | 214 (53.5%) | 298 (54.4%) |
| | Female | 186 (46.5%) | 251 (45.6%) |
| T stage [n (%)] | 1 | 9 (2.3%) | 16 (2.9%) |
| | 2 | 69 (17.2%) | 47 (8.5%) |
| | 3 | 275 (68.7%) | 372 (67.7%) |
| | 4 | 47 (11.8%) | 115 (20.9%) |
| Lymph node metastatic [n (%)] | Yes | 166 (41.5%) | 241 (43.8%) |
| | No | 234 (58.5%) | 309 (56.2%) |
| Distant metastatic [n (%)] | Yes | 57 (14.3%) | 60 (10.9%) |
| | No | 343 (85.7%) | 550 (89.1%) |
| TNM stage [n (%)] | I | 70 (17.5%) | 40 (7.2%) |
| | II | 156 (39.0%) | 257 (46.8%) |
| | III | 117 (29.3%) | 194 (35.3%) |
| | IV | 57 (14.2%) | 59 (10.7%) |
| Vital status | Alive | 318 (85.1%) | 374 (68%) |
| | Dead | 82 (14.9%) | 176 (32%) |
| Median follow up (months) | 23 | 53 |
FIGURE 3 | Identification of prognostic genes and nine-gene risk model construction in the TCGA cohort. (A) LASSO coefficients. (B) Plots of the cross-validation error rates. The dashes signify the value of the minimal error and greater λ value. (C) Multivariate Cox proportional hazard model of nine genes. (D) Risk score distribution in the two groups. (E) Survival overview in the two groups. (F) Heatmap of nine genes in the two groups. (G) Survival curve of the two groups. (H) Time-dependent ROC curve for 1-, 3-, and 5-years survival prediction. *p < 0.05, **p < 0.01, ***p < 0.001.
FIGURE 4 | Kaplan–Meier survival curves. Validation of the nine-gene model based on different clinicopathologic characteristics.
the cut-off criteria (|MM|>0.8, |GS|>0.3), twelve hub genes (LZTS1, TIE1, STARD8, VEGFC, KCNE4, ADAMTS1, AFAP1L1, ITGA5, BCL6B, MMRN2, CAVIN1 and CCBE1) with high connectivity in the clinical significant module were identified (Figure 8B). The expression of all twelve genes showed significant correlation with risk score (Supplementary Figures S3A–L). The hub genes were then analyzed in the GEPIA database. For LZTS1, VEGFC, KCNE4, ITGA5, CCBE1 and CAVIN1, K-M survival analysis revealed that a higher expression was meaningfully associated with a worse prognosis (Figures 8C–H).

Besides, GSEA was performed in the two risk groups. A false discovery rate (FDR) less than 0.25 and an absolute value of the normalized enrichment score (NES) greater than 1 were defined as the cut-off criteria. The gene sets of the high-risk group were enriched in tumor progression and metastasis-related pathways and inflammatory response-related pathways (Figure 9A), while the gene sets of the low-risk group were enriched in DNA integration, epithelial structure maintenance related pathways (Figure 9B).

Next, we analyzed gene mutations to further understand different molecular subgroup associated with the nine-gene risk score (Figures 9C,D). The mutation rates of APC, TP53, TTN, KRAS, SYNE1, MUC16 and PIK3CA were higher than 20% in both groups. The mutation of the RYR2, ABCA13 and PCL3 genes was more common in the high-risk group, while the mutation of FAT4, OBSCN and ZFHX4 genes was more common in the low-risk group.

**Immune Characteristics of Different Risk Groups**

Given that the features of tumor-infiltrating immune cells (TIICs) are correlated with the development and progression of cancer and may influence the prognosis, including CRC, we then used ssGSEA method to analyze the infiltration of various immune cells in tumor samples. We found that cells executing pro-tumor suppression (such as MDSCs, regulatory T cells, immature dendritic cells) were more abundant in the high-risk group (Figure 10A). Additionally, the correlation between the risk and common ICPs, i.e., PDCD1 (PD1), CD274 (PD-L1), CTLA4, LG3, HAVCR2 (TIM3) and TIGIT, was analyzed. Consistent with the higher infiltration of immunosuppressive cells in high-risk group, several common ICPs showed a significantly higher expression in the high-risk group (Figure 10B). The immune landscape and clinicopathological characteristics of different risk groups are presented as heat map in Figure 10C.

**DISCUSSION**

Pathological staging (tumor-node-metastases classification system) is a significant factor in clinical decision-making and prognosis evaluation of CRC. However, clinical outcomes often differ remarkably among patients at the same stage, suggesting that the current staging system is not sufficient in reflecting individual biological heterogeneity and predicting patient outcomes. A new prognostic assessment model referring to molecular and genetic profile may guide individualized therapy and improve long-term outcomes.

DNA methylation is an important mechanism in the regulation of gene expression. Aberrant methylation is frequently observed in tumors. These changes may promote malignant transformation by upregulating the expression of proto-oncogenes or inhibiting the expression of tumor suppressor genes. Abnormal methylation of some specific genes may be a key event in the early development of tumors and has potential to be a predictive biomarker for prognosis. With advances in sequencing technology, epigenetic changes of genes can be easily detected with a high degree of accuracy. Therefore, we adopted a beta mixture model-based method (MethylMix) to identify DNA methylation-driven genes. We identified 705 methylation-driven genes in CRC. Functional enrichment analysis revealed that these genes were involved in a wide range of biological processes and pathways, including cell signal transduction, cell differentiation, apoptosis regulation, metabolism, etc. These results suggest that DNA methylation is involved in gene dysregulation influencing various physiological processes and imply a possible mechanism by which DNA methylation is functionally associated with progression and prognosis in patients with CRC.

In our study, combining LASSO regression and multivariable Cox regression analysis, nine genes that were closely related to survival and prognosis were selected. Most of these genes have been reported in cancer studies. Glutathione S-transferase Mu 1 (GSTM1), a member of the glutathione S-transferase family, functions primarily as a detoxification enzyme, also involved...
FIGURE 5 | Nomogram for survival prediction. (A) Multivariable Cox proportional hazard model of the risk score and clinicopathologic factors. (B) OS-associated nomogram. (C) Time-dependent ROC curve for 1-, 3-, and 5-years survival prediction. **p < 0.01, ***p < 0.001.
in the negative regulation of apoptosis-related signaling pathways (Kalinina et al., 2014). Aberrant methylation of GSTM1 has been found in various cancers, such as acute myeloid leukaemia, urothelial carcinoma and head and neck cancer (Sharma et al., 2010; Ohgami et al., 2012; Wang et al., 2016). However, the methylation changes in CRC remain to be further validated. Heat

FIGURE 6 | Validation of the prediction model. (A) Survival curve of the two groups. (B–D) Nomogram calibration plots at 1-, 3-, 5-years in the validation cohort. (E) Time-dependent ROC curve for 1-, 3-, and 5-years in the validation cohort.
FIGURE 7 | Construction of weighted gene correlation network. (A, B) Screening and validation of the soft threshold. (C) Clustering dendrogram of genes. (D) Correlation between modules and risk model and identification of the key module.
FIGURE 8 | Analysis of weighted gene correlation network. (A) Enrichment analysis in the Metascape database and the top 20 enrichment terms were shown. (B) Identification of the hub genes (|MM| > 0.8, |GS| > 0.3). (C) Association of OS and LZTS1 expression in GEPIA database. (D) Association of OS and VEGFC expression in GEPIA database. (E) Association of OS and KCNE4 expression in GEPIA database. (F) Association of OS and ITGA5 expression in GEPIA database. (G) Association of OS and CCBE1 expression in GEPIA database. (H) Association of OS and CAVIN1 expression in GEPIA database.
FIGURE 9 | Molecular characteristics of different risk groups. (A) Gene sets enriched in high-risk group. (B) Gene sets enriched in low-risk group. (C) Significantly mutated genes in the mutated samples of high-risk group. (D) Significantly mutated genes in the mutated samples of low-risk group. Mutated genes (rows, top 10) are ordered by mutation rate. The right shows mutation percentage, and the top shows the overall number of mutations. The color-coding indicates the mutation type.
FIGURE 10 | Immune characteristics of different risk groups. (A) The relative abundance of TiICs in different risk groups. (B) The expression of ICPs in different risk groups. The scattered dots represent the outliers of the two risk groups. The thick lines represent the median value. The bottom and top of the boxes are the 25th and 75th percentiles (interquartile range), respectively. Significant statistical differences between the two groups were assessed using the Wilcoxon test. (C) The relative abundance of TiICs for 400 patients in the TCGA cohort. ns: not significant, *p < 0.05, **p < 0.01, ***p < 0.001.
hypomethylation is associated with platinum resistance in ovarian
non-small cell lung cancer (Ma et al., 2017), whereas the
mediate the gene inactivation (Mokarram et al., 2009).
PTPRD promoter has been validated in CRC, which may
methylation-specific (ARHGAP4) belongs to the small GTPase family, which can
hydrolyze the active GTP into inactive GDP and negatively
RhoA protein. Studies showed that it was involved in
regulating the invasion and metastasis of pancreatic cancer cells
(Shen et al., 2019a; Shen et al., 2019b). Receptor-type tyrosine-
protein phosphatase delta (PTPRD) is frequently found to be
inactivated by epigenetic modification in a variety of tumors, suggesting that it may have tumor suppressive effects (Kohno et al., 2010; Funato et al., 2011; Gieffing et al., 2011). Based on
methylation-specific PCR, frequently hypermethylation of
PTPRD promoter has been validated in CRC, which may
mediate the gene inactivation (Mokarram et al., 2009).
Furthermore, in vitro experiments also showed that exogenous
expression of PTPRD could inhibit the migration and invasion of
colon cancer cells (Funato et al., 2011). Transmembrane protein
88 (TMEM88), a newly discovered protein that is localized on cell
membranes, can inhibit the classical Wnt signaling pathway and
is thought to have a bidirectional effect of inhibiting or promoting
cancer in different contexts (Ge et al., 2018). Indeed, promoter
hypermethylation of TMEM88 is associated with poorer prognosis of
non-small cell lung cancer (Ma et al., 2017), whereas the
hypermethylation is associated with platinum resistance in ovarian
cancer (de Leon et al., 2016). What effect of aberrant TMEM88
methylation has on CRC needs to be further explored. LINC01555 is a
long non-coding RNA that has rarely been studied. Von Willebrand factor D and EGF domain-containing protein (VWDE) is a blastema-
enriched gene in a variety of highly regenerative species (Leigh et al.,
2020). A recent study showed that VWDE is a significant driver
driving oncogene with a highly mutation prevalence in breast cancer (Pongor et al., 2015). Abhydrolase domain containing 12B (ABHD12B) locates on chromosome 14. Physiological and pathological functions of
ABHD12B are rarely reported. Given the significant association of
these three genes (LINC01555, VWDE and ABHD12B) with prognostic in CRC, it is meaningful to further study that how they
regulate the development of CRC.

Furthermore, based on the expression of these nine genes and coefficients with survival, a prognostic model was
established. Each patient’s risk score was calculated and all
patients were divided into two groups based on the X-tile
software. The K-M survival curve revealed that the model
could distinguish the difference of OS between the two risk
groups. Time-dependent ROC analysis showed that the nine-
gene risk model had a good performance in predicting OS of
CRC. Our risk model was a promising prognostic biomarker for CRC patients.

To determine the molecular mechanisms which drive the nine
gene risk model, we took many approaches. Firstly, WGCNA was
performed to explore clinical significant modules associated with
the nine-gene signature. The blue module, which contained 1,577
genes, was found to have the highest association with risk model.
GO/KEGG enrichment analysis by Metascape showed that the functions of “extracellular matrix organization,”
“vasculature development,” “cell junction organization,” “cell-substrate
adhesion,” “collagen fibril organization” were significantly
enriched. Many of these pathways have been reported to be
involved in tumor onset and development. According to the
cut-off criteria |MM| > 0.8, |GS| > 0.3, twelve genes with a high
connectivity were screened out from the clinical significant
module. The analysis using GEPIA also confirmed that high
expression of LZTS1, VEGFC, KCNE4, ITGA5, CBE1 and
Cavin1 was significantly associated with poorer prognosis in
CRC. Leucine zipper putative tumor suppressor 1 (LZTS1) was
reported to suppress cancer cell growth and regulate cell cycle
(Ishii et al., 2001). However, given that the higher expression was
associated with shorter OS, the specific role of LZTS1 in CRC
needs further research. Vascular endothelial growth factor C
(VEGFC) is active in blood vessels and lymphatics endothelial
cell proliferation and migration, therefore contributing to tumor
metastases (Tacconi et al., 2015). Recently, VEGFR3 was found to
be expressed in tumor-associated macrophages (TAMs) in CRC.
VEGFC/VEGFR3 pathway could inhibit antitumor immunity by
regulating TAMs (Tacconi et al., 2019). Potassium voltage-gated
channel subfamily E member 4 (KCNE4) has been found to be
increased in multiple solid tumors (Biasiotta et al., 2016). Some
studies also show a direct association between the transmembrane ion transport and carcinogenesis, although the
exact mechanism is still unclear (Cone, 1969). Integrin alpha-5
(ITGA5) can form heterodimer with different beta subunits.
Down-regulation of ITGA5 in CRC cells could inhibit cell
proliferation and tumorigenesis and promote cell apoptosis
(Yu et al., 2019). Besides, one study shows that ITGA5 is
required for the tumor supportive role of fibroblasts (Lu et al.,
2019). Collagen and calcium-binding EGF domain-containing
protein 1 (CBE1) is important for lymphatic vascular
development and plays a pro-tumor role in colorectal cancer
by promoting lymphangiogenesis and lymphatic metastasis of
cancer cells (Song et al., 2020). Indeed, in the TCGA cohort
CBE1 expression was significantly higher in patients with
lymph node metastasis compared to those without lymph
node metastasis (Supplementary Figure S4A). Caveola-
associated protein 1 (Cavin1) cooperates with Caveolin to
regulate lipid uptake of cell. Cavin1 can promote the secretion
of extracellular vesicles (EVs) in glioma, and EVs expressing
cavin1 in turn promote the growth of glioma (Wang et al., 2020).

We also used GSEA to analyze the differential gene expression
between the high-risk group and the low-risk group. Consistent
with the results of WGCNA, the biological processes or pathways
enriched in the high-risk group were mostly involved in tumor
invasion and progression. For example, activation of the
transcription factor STAT3 has been widely reported in
various tumors, including CRC (Bollrath et al., 2009). Phosphorylation and nuclear translocation of STAT3 drive transcription of genes involved in cell-cycle regulation, cell survival, cell migration, ultimately promoting invasion and metastasis of tumors (Yu et al., 2014).

Numerous studies have revealed that the features of tumor-infiltrating immune cells (TIICs) are closely correlated with the development and progression of cancer (Grivennikov et al., 2010; Steidl et al., 2010; Herrera et al., 2013; Waniczek et al., 2017; Ye et al., 2018). TAMs as a prognostic risk factor for several kinds of cancer, including CRC (Steidl et al., 2010; Herrera et al., 2013; Waniczek et al., 2017; Ye et al., 2019). Accordingly, the expression of several common ICPs was higher in the high-risk group, further revealing that the high-risk group was characteristics of immunosuppression.

The efficacy of every single biomarker is inadequate. To further enhance the potential clinical application of this risk model, we established a nomogram. Nomogram combined multiple risk factors (nine-gene signature, age, T stage and distant metastatic) and constructed a simple statistical prediction model. Time-dependent ROC and calibration curve both showed that the predicted survival rate was close to the actual situation. Our nomogram might serve as an excellent predictive model in the prognosis assessment of CRC patients.

To the best of our knowledge, this nine-gene risk model has not been previously reported, and the information of DNA methylation driven genes provides a foundation for further exploration of the pathogenesis and therapeutic strategy of CRC. However, there are some limitations in current study. The risk model requires further confirmation by increasing the sample size and performing prospective, multicenter studies.

**REFERENCES**

Anders, S., and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. * Genome Biol.* 11 (10), R106. doi:10.1186/gb-2010-11-10-r106

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global Patterns and Trends in Colorectal Cancer Incidence and Mortality. * Gut* 66 (4), 683–691. doi:10.1136/gutjnl-2015-310912

Biaisotta, A., D’Arcangelo, D., Passarelli, F., Nicodemi, E. M., and Facchiano, A. (2016). Ion Channels Expression and Function Are Strongly Modified in Solid Tumors and Vascular Malformations. * J. Transl. Med.* 14 (1), 285. doi:10.1186/s12967-016-1038-y

Bühl, M. P., Foerster, A., Lugli, A., and Zlobec, I. (2012). Characterization of CDKN2A(p16) Methylation and Impact in Colorectal Cancer: Systematic Analysis Using Pyrosequencing. * J. Transl. Med.* 10, 173. doi:10.1186/1479-5876-10-173

Bollrath, J., Phesse, T. J., von Burstin, V. A., Putoczki, T., Bennecke, M., Bateman, T., et al. (2009). gp130-mediated Stat3 Activation in Enterocytes Regulates Cell Survival and Cell-Cycle Progression during Colitis-Associated Tumorigenesis. * Cancer Cell* 15 (2), 91–102. doi:10.1016/j.ccr.2009.01.002

**CONCLUSION**

In summary, our study established a nomogram that combined the DNA methylation-driven genes and some significant clinicopathological features. This model has been validated in different study cohorts and is expected to serve as a predictive tool with high sensitivity and specificity to assess clinical prognosis and to guide individualized anti-tumor therapy for patients with CRC.

**DATA AVAILABILITY STATEMENT**

Publicly available datasets were analyzed in this study. This data can be found here: GSE39582.

**AUTHOR CONTRIBUTIONS**

WW designed this study. ZF and WW collected the data, analyzed the data and wrote the manuscript. ZL and PK revised the manuscript. All authors approved the final version for submission.

**FUNDING**

This work was supported by grants from the National Natural Science Foundation of China (82070562, 81630017, and 91942312), Shanghai Science and Technology Development Funds (20QA1407700).

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.779383/full#supplementary-material
Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a Web Server for Cancer and normal Gene Expression Profiling and Interactive Analyses. Nucleic Acids Res. 45 (W1), W98–w102. doi:10.1093/nar/gkw247

Wang, L., Yang, C., Wang, Q., Liu, Q., Wang, Y., Zhou, J., et al. (2020). Homotrimer Cavin1 Interacts with Caveolin1 to Facilitate Tumor Growth and Activate Microglia through Extracellular Vesicles in Glioma. Theranostics 10 (15), 6674–6694. doi:10.7150/thno.45688

Wang, S.-C., Huang, C.-C., Shen, C.-H., Lin, L.-C., Zhao, P.-W., Chen, S.-Y., et al. (2016). Gene Expression and DNA Methylation Status of Glutathione S-Transferase Mu1 and Mu5 in Urothelial Carcinoma. PloS one 11 (7), e0159102. doi:10.1371/journal.pone.0159102

Waniczek, D., Lorenc, Z., Sniertura, M., Wesecki, M., Kopec, A., and Mucz-Wierzgosn, M. (2017). Tumor-Associated Macrophages and Regulatory T Cells Infiltration and the Clinical Outcome in Colorectal Cancer. Arch. Immunol. Ther. Exp. 65 (5), 445–454. doi:10.1007/s00005-017-0463-9

Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., and Ji, Y. (2018). TCGA-assembler 2: Software Pipeline for Retrieval and Processing of TCGA/CPTAC Data. Bioinformatics 34 (9), 1615–1617. doi:10.1093/bioinformatics/btx812

Yanagi, T., Nagai, K., Shimizu, H., and Matsuzawa, S.-I. (2017). Melanoma Antigen A12 Regulates Cell Cycle via Tumor Suppressor P21 Expression. Oncotarget 8 (40), 68448–68459. doi:10.18632/oncotarget.19497

Ye, L., Zhang, T., Kang, Z., Guo, G., Sun, Y., Lin, K., et al. (2019). Tumor-Infiltrating Immune Cells Act as a Marker for Prognosis in Colorectal Cancer. Front. Immunol. 10, 2368. doi:10.3389/fimmu.2019.02368

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A J. Integr. Biol. 16 (5), 284–287. doi:10.1089/omi.2011.0118

Yu, H., Lee, H., Herrmann, A., Buettner, R., and Jove, R. (2014). Revisiting STAT3 Signalling in Cancer: New and Unexpected Biological Functions. Nat. Rev. Cancer 14 (11), 736–746. doi:10.1038/nrc3818

Yu, M., Chu, S., Fei, B., Fang, X., and Liu, Z. (2019). O-GlcNAcylation of ITGAV5 Facilitates the Occurrence and Development of Colorectal Cancer. Exp. Cell Res. 382 (2), 111464. doi:10.1016/j.yexcr.2019.06.009

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. Nat. Commun. 10 (1), 1523. doi:10.1038/s41467-019-09234-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Feng, Liu, Peng and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.