Identification of methylation driven biomarkers for diagnosis and prognosis in colorectal cancer by Integrative Analysis of TCGA, GTEx, and GEO database

Lichao Cao

Northwest University

Tong Li

Northwest University

Ying Ba

Shenzhen Nucleus Gene Technology Co., Ltd

Erfei Chen

Northwest University

Jin Yang (✉ yangjin@nwu.edu.cn)

Northwest University

Hezi Zhang

Shenzhen Nucleus Gene Technology Co., Ltd

Research Article

Keywords:

Posted Date: February 28th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-964805/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: This work investigates the use of methylation driven biomarkers for diagnosis and prognosis in colorectal cancer (CRC) by mining DNA methylation and gene expression data from The Cancer Genome Atlas (TCGA), the Genotype-Tissue Expression project (GTEx), and the Gene Expression Omnibus (GEO). **Methods:** The differentially expressed genes (DEGs) and differentially methylated genes (DMGs) were screened using mRNA expression and DNA methylation data from TCGA, respectively. The methylation driven genes (MDGs) of CRC were further identified using the MethylMix R package. Subsequently, the MDGs were analyzed with Random Forest (RF), support vector machine (SVM), and logistic regression (LR) algorithms to establish diagnosis prediction models as independent indicators using mRNA expression data from TCGA and GTEx. The RF algorithm was determined to be the most suitable and used to construct the diagnostic model with the combined MDGs, which was then validated by GSE39582 from GEO. Prognostic biomarkers were used to establish the risk score model, which was generated by univariate and multivariate Cox regression analyses. Moreover, we constructed and validated a nomogram that integrated the risk score and clinical information, including age, gender, and tumor stage. **Results:** 9 out of 10 MDGs performed well as independent diagnostic predictors, and STK33 and EPHX4 were also found to be associated with overall survival (OS). The results of the nomogram suggest that it is a better predictive model for prognosis than the risk score model. **Conclusion:** Our findings suggest that the identified MDGs could be biomarkers for diagnosis and prognosis of CRC.

Introduction

CRC is one of the most prevalent malignant tumors and the second leading cause of mortality worldwide[1]. Despite the progress made in diagnosis and therapy, CRC patients usually develop recurrence and metastasis, leading to dramatic decreases in the 5-year survival rate[2]. Therefore, there is an urgent need to improve the diagnosis, treatment, and prognosis for patients with CRC.

Molecular characterization has great potential for improving understanding of tumor development and is extensively utilized to predict tumor diagnosis, treatment, and prognosis[3, 4]. Using bioinformatics methods and machine learning, various types of biomarkers have been found to be related to the diagnosis and prognosis of tumors, including microRNAs[5, 6], long non-coding RNAs[7, 8], DEGs[9], DNA methylation [10, 11] and others.

DNA methylation as an important regulator of gene expression has been researched in various cancers, such as endometrial cancer[12], prostate cancer [13](R Lesche et al. 2007), and hepatocellular carcinoma [14](RH Xu et al. 2017). Although the mechanism of DNA methylation is not fully understood, it is assumed that DNA methylation could affect the binding of the transcription factors to their DNA targets sites, and then alter the expression of downstream genes [15-17](R Jaenisch & A Bird 2003; T Vaissiere et al. 2008; M Yassi et al. 2018), which may promote oncogenesis [15, 18](G Egger et al. 2004; JG Herman & SB Baylin 2003). MethylMix, a R package that can identify MDGs through integrative analysis of DNA methylation and gene expression data from normal samples and tumor samples[19, 20], was used to

investigate the relationship between MDGs and the prognosis of CRC patients[21, 22]. However, no studies have reported whether MDGs could be used as diagnostic indicators for CRC.

The public databases, especially TCGA and GEO, provide convenient access to systematic collections of sequencing data with detailed clinical and pathological information which have been applied in malignant tumor research. Meanwhile, the GTEx database contains a large number of gene expression samples of normal human patients that have been integrated with TCGA or GEO in various studies [23-25].

Based on existing literature data, we carried out studies to identify a set of MDGs using DNA methylation and gene expression data from TCGA, GEO, and the GTEx, and construct an independent diagnosis model using RF, SVM, and LR algorithms. Additionally, the MDGs, the presumed potential prognostic indicators, were analyzed using univariate and multivariate Cox regression analyses. Our findings may suggest that the potential methylation driven biomarkers could prompt more individualized diagnoses and therapies for CRC.

Materials And Methods

2.1 TCGA and GTEx Data acquisition

The mRNA expression (RNA-seq data from Illumina platform) and DNA methylation data (Illumina Human Methylation 450) of CRC from TCGA and the mRNA expression of normal samples from the GTEx were downloaded from the UCSC Xena platform (<https://xenabrowser.net/datapages/>). Subsequently, the samples from TCGA were divided into the normal group and the tumor group. 667 samples (51 normal and 616 tumor) were selected for DEG analysis and MDG analysis was performed on 433 samples (45 normal and 388 tumor), detailed information is shown in Table 1. We selected 307 normal samples (123 females and 184 males) of mRNA expression data of colon tissue from the GTEx to provide supplementary data for normal samples from TCGA.

2.2 GEO data acquisition

We downloaded gene expression data by array (GSE21815, GSE28000, GSE39582, and GSE44076) and methylation data by genome tiling array (GSE48684, GSE53051, GSE77718, and GSE101764) from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>), and the sample size of each dataset can be found in Figure 1. Among them, GSE39582 has detailed clinical and phenotypic information, with a total of 566 tumor and 19 normal samples (Table 1).

2.3 Screening DEGs, DMGs and MDGs

We filtered out the genes with a read count of less than 10 in more than 50% of all the samples. The DEGs were identified by comparing the tumor group with the normal groups using the limma R package[26], with the cutoff criteria defined as $|\log_2 \text{fold change (FC)}| > 0.585$ and *adjusted P-value* < 0.001. In addition, the probe with the highest absolute value of *deltaBeta* and *adjusted P-value* less than

0.001 was selected as the representative of gene methylation level using the ChAMP R package[27], and $|\Delta\beta| > 0.3$ was used to screen DMGs.

Subsequently, the tumor methylation matrix, the normal methylation matrix, and the tumor expression matrix with overlapping DEGs and DMGs were constructed to identify MDGs using the R package MethylMix[19, 20]. Significant methylation events were filtered using the *correlation coefficient* < -0.3 and *P-value* < 0.05 as the cutoff criteria, and the correlation of the MDGs was visualized by the corplot R package. The unpaired t test statistically analyzed the differences of the MDGs between the tumor samples and the normal samples.

2.4 Validation of MDGs using GEO datasets

To validate the MDGs, the gene expression profiling datasets (GSE21815, GSE28000, GSE39582, and GSE44076) were separately analyzed by the online software GEO2R (<http://www.ncbi.nlm.nih.gov/geo/geo2r/>), and the expression value of each MDG was obtained. The methylation value of each MDG in the gene methylation profiling datasets (GSE48684, GSE53051, GSE77718, GSE101764) was calculated using the ChAMP R package.

2.5 Identification and validation of diagnosis biomarkers

The mRNA expression datasets from TCGA and the GTEx were integrated and randomly divided into a training dataset and a validation dataset with a 3:1 ratio in tumor samples and normal samples; principal component analysis (PCA) investigated the distribution of the combined samples of TCGA and the GTEx. To choose the optimal algorithm to build the diagnosis model, three different algorithms were used to evaluate the accuracy of the 10 MDGs as an independent predictor with default parameters and calculate the *P-value* of each MDG. The chosen algorithm was used to further evaluate the combined indicator for diagnosis of CRC. The GSE39582 dataset, containing 7 of the MDGs, was analyzed as a validation dataset in combination with training datasets from TCGA and the GTEx, the receiver operating characteristic (ROC) curve was utilized to depict the sensitivity and specificity of the diagnosis models using independent MDGs or the combined MDGs panel.

2.6 Identification and evaluation of prognosis biomarkers

The mRNA expression data with OS information was selected to determine the prognostic biomarkers. The univariable Cox proportional hazards regression model was produced using the Survival R package, with the *P-value* cutoff of 0.05. Subsequently, Kaplan-Meier survival and multivariate Cox regression analyses were performed to determine independent prognostic factors; the coefficient of the MDGs was obtained from the multivariate Cox results. The risk score based on the signature of each CRC patient was calculated using the following formula:

MDGs risk score=

To further evaluate the predictive efficiency of the constructed MDGs risk score model, we used the ROC curve to reflect the sensitivity and specificity of survival prediction and quantified the area under the curve (AUC) using the survivalROC R package. The optimal cutoff risk score was designated at the turning point of the ROC curve, where the difference between true positive and false positive is the most significant. The patients above the cutoff value were in the high-risk group, while the patients below it was in the low-risk group. In addition, Kaplan-Meier curves were plotted to distinguish the two groups using the survminer R package.

2.7 Construction and validation of the nomogram model

To improve the accuracy of the prognostic model, we constructed a nomogram that visualized the prognostic value of different patients' characteristics by integrating risk score and clinical information, including age, gender, and tumor stage. This analysis was performed using the *rms* R package to plot calibration curves to evaluate the predicted probabilities in comparison with the ideal predictive line. In addition, a forest plot based on univariable Cox analysis illustrated the relationship between clinical information and OS. Moreover, the Concordance index (C-index) indicated the predictive accuracy of the nomogram.

Results

3.1 Identification of MDGs as candidate diagnostic and prognostic biomarkers

The data generation and analysis workflow are shown in Figure 1. After merging the expression datasets and methylation datasets of colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) from TCGA, and expression datasets from TCGA and the GTEx, respectively, PCA showed that the normal samples and the tumor samples were well separated. The samples of READ and COAD, TCGA, and the GTEx were randomly distributed, which indicated that the merged data sets could be used for further analysis (Figure S1 A-B). Expression and DNA methylation data from TCGA were separately analyzed to screen DEGs and DMGs by comparing normal samples and tumor samples. As a result, 522 DEGs (266 upregulated and 256 downregulated) were identified (Figure 2A and Additional file 1: Table S1), and the detailed information about DMGs was shown in Additional file2: Table S2.

Subsequently, R package MethylMix was used for identifying MDGs with a filter criterion set as $Cor < -0.3$, $| \Delta\beta | > 0.3$ and $adjusted\ P\text{-value} < 0.001$, and 10 MDGs were identified as candidate diagnostic and prognostic biomarkers. Pearson's correlation test statistically analyzed the correlations between methylation degree and gene expression of the MDGs using the cor.test function of the R language (<https://www.r-project.org/>) (Figure 3) and found that the methylation degree of those 10 MDGs was negatively correlated with their expression. The higher the methylation degree was, the lower the gene expression, and the detailed information of the MDGs is shown in Table 2. The correlations between each candidate biomarker, which may help us choose the optimal independent gene set for the diagnosis of CRC, were investigated using the corrplot R package (Figure 2B). In addition, the unpaired *t* test was

conducted to quantify the difference of each candidate biomarker between the normal samples and the tumor samples; significant difference was found in all the candidate biomarkers (Figure 2C)

Moreover, we verified the gene differences and methylation differences of these MDGs in other datasets (4 expression GEO datasets and 4 methylation GEO datasets) and found that the trend was consistent. The significant differences of gene expression and methylation of these MDGs were verified in at least one other dataset (Table 3). The DEGs of each GEO dataset is shown in Figure S1 C-F.

3.2 Evaluation of the MDGs as candidate diagnostic biomarkers

We used three different algorithms to evaluate the performance of MDGs in establishing prognosis models. Among them, SVM algorithms performed the worst (Figure S2); RF and LR algorithms had similar results, which can be seen in Figure 4 and Figure S3, respectively. However, when using the LR algorithm to construct prediction models with joint MDGs, the algorithm does not aggregate; using the RF algorithm, the calculated *P-value* of each MDG is less than 0.0001, while in the other two algorithms, the *P-value* of *SPTBN5* is greater than 0.0001 (Table 4). Therefore, the RF algorithm was selected to further build the diagnostic model of the joint MDGs. When using MDGs to build prediction models with the RF algorithm, 9 out of 10 MDGs revealed excellent performance as independent diagnostic predictors, where the area under the curve (AUC) of *CLDN1* is 0.988, *EPHX4* is 0.971, *TCN1* is 0.966, *ARHGAP20* is 0.921, *LY6G6D* is 0.886, *FAM150A* is 0.869, *KCNJ12* is 0.857, *KRT7* is 0.860, and *STK33* is 0.842 (Figure 4 A-I). Additionally, the performance of the combined MDGs were further assessed in samples from TCGA-GTEx and GEO and the outcome of the AUC was 1.0 and 0.996, respectively (Figure 4 K, L). The importance of each predictor for the combined predictive model is shown in Table 5. The larger the number of *MeanDecreaseAccuracy* and *MeanDecreaseGini*, the more important the predictor for the model. It was found that the numbers of *MeanDecreaseAccuracy* and *MeanDecreaseGini* were positively correlated with the AUC of the predictive model of the predictor. Moreover, unsupervised hierarchical clustering of the combined panels with ten makers or seven makers indicated that the constructed models could accurately distinguish CRC patients from normal controls (Figure 2D-G). Collectively, the 10 MDGs could be candidate biomarkers for diagnosis of CRC samples and 9 of them performed well as independent indicators.

3.3 Construction and evaluation of MDGs related prognostic model in CRC patients

Univariate Cox regression analysis revealed that 2 of the 10 MDGs were independent prognostic indicators of OS, as the hazard ratio of *STK33* was 1.09 (95 % *CI*: 1.01-1.17, *P* = 0.021) and *EPHX4* was 0.91 (*CI*: 0.83-0.99, *P* = 0.032) (Figure 5A). Subsequently, Kaplan-Meier analyses and log-rank tests using *STK33* and *EPHX4* as independent prognostic indicators indicated that patients with high-risk scores suffered poor OS, with the *P-value* of 0.008 and 0.047, respectively (Figure 5B, C). We calculated the risk score of each CRC patient using the formula as follows: (1.09 *STK33*) + (0.91 *EPHX4*); then CRC patients (n=635) were categorized into the high-risk score group or the low-risk score group, according to the optimal cut-off value of the risk score obtained from the survminer R package. The results also indicated that high-risk score patients had a worse OS rate than low-risk score patients (*P* < 0.0001)

(Figure 5D). The prognostic accuracy of the risk score model was investigated as a continuous variable (Figure 5E). The AUC of the prognostic model for OS was 0.569 at 3 years, 0.633 at 4 years, and 0.626 at 5 years.

3.4 Construction and validation of the nomogram model

In the nomogram the score for each variable can be found on the point scale, then used to estimate the probability of survival at 1, 3, and 5 years by calculating the total score (Figure 6A). The forest plot shows that patient characteristics, including age (>60), tumor stage (III and IV), and risk score are associated with OS (P -value <0.001) (Figure 6B).

To validate the nomogram's performance, we plotted the calibration curves and observed that the predictive curves were close to the ideal curve (Figure 6C-E), which indicates good performance. Furthermore, the predictive accuracy of this nomogram (C-index: 0.72) was higher than the risk score model (C-index: 0.57).

Discussion

Many previous studies have reported that DMGs and DEGs can be used as prognostic and diagnostic biomarkers for CRC[11, 28-30]. A previous study also reported that MDGs can assist with determining the prognosis of CRC patients[21]. However, no studies researched the relationship between MDGs and the diagnoses of CRC patients. In this study, we investigated the interplay between DMGs and DEGs and chose the genes with negative and high correlation, namely MDGs, as candidate predictors for the prognosis and diagnosis of CRC. We identified 10 MDGs that could be considered as potential diagnostic and prognostic predictors for CRC patients: *CLDN1*, *EPHX4*, *TCN1*, *LY6G6D*, *FAM150A*, *KRT7*, *STK33*, *ARHGAP20*, *KCNJ12* and *SPTBN5*. Among them, *STK33* and *EPHX4* were correlated with both diagnosis and prognosis of colorectal cancer, while the others only performed well as diagnostic indicators of colorectal cancer. Consistent with our findings, various studies have demonstrated that *STK33* is overexpressed in hypopharyngeal squamous cell carcinoma [31], hepatocellular carcinoma[32], and human large cell lung cancer[33]. *STK33* hypermethylation may be a biomarker for the diagnosis, prognosis, and suitable treatment of CRC [34](MD Yin et al. 2018). Previous studies demonstrated that *EPHX4* was significantly upregulated in rectal cancer [35](H Flebbe et al. 2019) and pseudomyxoma peritonei [36]. However, the pathological role and clinical significance of *EPHX4* for CRC have rarely been reported. As a member of the claudin family and tight junction-related proteins, *CLDN1* was demonstrated to be associated with dysfunction or abnormal expression in various tumors [37, 38]. In addition, a previous study revealed that aberrant expression of *CLDN1* regulated the *AMPK/STAT1/ULK1* signaling pathways, leading to the promotion of proliferation and metastasis in esophageal squamous cancer[39]. *CLDN1* was experimentally demonstrated to be remarkably upregulated in CRC patients and could be considered as a methylated diagnostic biomarker in CRC patients and normal control groups[40]. A growing number of studies have verified that the overexpression of *TCN1* is associated with tumor invasion and metastasis in CRC [41, 42] and experimentally demonstrated that *TCN1* was significantly

overexpressed in colon cancer tissues compared with normal controls at the mRNA and protein level, and could be considered as potential prognostic biomarker of colon cancer[43]. A previous study showed that *KRT7* plays a significant role in tumor metastasis and is considered as a prognostic biomarker and potential target for therapeutic prevention of metastasis[44](L Jiang et al. 2019). In addition, *KRT7* was down-regulated and hypermethylated in CRC tissues compared with adjacent normal tissues and may lead to the occurrence of CRC[45](JY Wang et al. 2017). *Y6G6D* belongs to a cluster of leukocyte antigen-6 genes, which was conspicuously overexpressed (around 15-fold) and considered a promising biomarker of immunotherapy for microsatellite stable CRC[46]. *FAM150A* can activate lymphoma kinase (*ALK*) by binding to its extracellular domain[47, 48]; *ALK* has been used as an effective biomarker in various human cancers, such as neuroblastoma and non-small cell lung cancer [49]. Additionally, the DNA methylation status of *FAM150A* was indicated to be a diagnostic and prognostic indicator for clear cell renal cell carcinoma (ccRCC) using the high-performance liquid chromatography method[50]. However, no studies found that *FAM150A* was associated with CRC. To the best of our knowledge, there is still a lack of research about the relationship between *ARHGAP20*, *KCNJ12*, and *SPTBN5* and oncogenesis, which may represent novel predictive biomarkers.

Furthermore, we constructed a nomogram based on the multivariable Cox regression coefficients of risk score, age, gender, and tumor stage to further validate our findings. The C-index of the nomogram was significantly higher than the C-index of the risk score model, which was remarkably associated with OS, suggesting that the two MDGs could be used as prognostic indicators.

Conclusions

In conclusion, we identified 10 MDGs that could be used as potential biomarkers for CRC, of which 9 performed well as independent diagnostic predictors and 2 could be used as prognostic indicators.

Declarations

SUPPORTING INFORMATION

Additional file 1: Table S1. Detailed information of DEGs.

Additional file 2: Table S2. Detailed information of DMGs

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in TCGA and GTEx from the UCSC Xena (<https://xenabrowser.net/datapages/>) and GEO with accession numbers GSE21815, GSE28000, GSE39582, GSE44076, GSE48684, GSE53051, GSE77718, and GSE101764.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Key Science and Technology Program of Shaanxi Province (2019ZDLSF02-05) and National Natural Scientific Foundation of China (81974378 and 82003115).

AUTHOR CONTRIBUTIONS

Lichao Cao and Erfei Chen were involved in the study concept and design, and drafting the manuscript. Hezi Zhang and Ying Ba put forward some kind suggestions. Tong Li and Jin Yang helped analyze and interpret the data.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

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: a cancer journal for clinicians* 2018, **68**(6):394-424.
2. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, Cercek A, Smith RA, Jemal A: **Colorectal cancer statistics, 2020.** *CA: a cancer journal for clinicians* 2020, **70**(3):145-164.
3. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM: **Identification and characterization of essential genes in the human genome.** *Science* 2015, **350**(6264):1096-1101.
4. Dexheimer GM, Alves J, Reckziegel L, Lazzaretti G, Abujamra AL: **DNA Methylation Events as Markers for Diagnosis and Management of Acute Myeloid Leukemia and Myelodysplastic Syndrome.** *Disease markers* 2017, **2017**:5472893.
5. Chen W, Gao C, Liu Y, Wen Y, Hong X, Huang Z: **Bioinformatics Analysis of Prognostic miRNA Signature and Potential Critical Genes in Colon Cancer.** *Front Genet* 2020, **11**:478.
6. Zhou XG, Huang XL, Liang SY, Tang SM, Wu SK, Huang TT, Mo ZN, Wang QY: **Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis.** *OncoTargets and therapy* 2018, **11**:2815-2830.
7. Lin Y, Pan X, Chen Z, Lin S, Chen S: **Identification of an Immune-Related Nine-lncRNA Signature Predictive of Overall Survival in Colon Cancer.** *Front Genet* 2020, **11**:318.
8. Jin L, Li C, Liu T, Wang L: **A potential prognostic prediction model of colon adenocarcinoma with recurrence based on prognostic lncRNA signatures.** *Human genomics* 2020, **14**(1):24.
9. Yang Z, Chen Y, Wu D, Min Z, Quan Y: **Analysis of risk factors for colon cancer progression.** *OncoTargets and therapy* 2019, **12**:3991-4000.

10. Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, Flagg K, Hou J, Zhang H, Yi S *et al*: **DNA methylation markers for diagnosis and prognosis of common cancers.** *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114**(28):7414-7419.
11. Yang C, Zhang Y, Xu X, Li W: **Molecular subtypes based on DNA methylation predict prognosis in colon adenocarcinoma patients.** *Aging* 2019, **11**(24):11880-11892.
12. Ying J, Xu T, Wang Q, Ye J, Lyu J: **Exploration of DNA methylation markers for diagnosis and prognosis of patients with endometrial cancer.** *Epigenetics-Uk* 2018, **13**(5):490-504.
13. Lesche R, Payne S, Model F, Weiss G, Sledziewski A: **DNA methylation markers for diagnosis and prognosis of prostate cancer.** *Tumor Biol* 2007, **28**:51-51.
14. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, Yi S, Shi W, Quan Q, Li K *et al*: **Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma.** *Nature materials* 2017, **16**(11):1155-1161.
15. Herman JG, Baylin SB: **Gene silencing in cancer in association with promoter hypermethylation.** *The New England journal of medicine* 2003, **349**(21):2042-2054.
16. Vaissiere T, Sawan C, Herceg Z: **Epigenetic interplay between histone modifications and DNA methylation in gene silencing.** *Mutation research* 2008, **659**(1-2):40-48.
17. Yassi M, Shams Davodly E, Mojtabanezhad Shariatpanahi A, Heidari M, Dayyani M, Heravi-Moussavi A, Moattar MH, Kerachian MA: **DMRFusion: A differentially methylated region detection tool based on the ranked fusion method.** *Genomics* 2018, **110**(6):366-374.
18. Egger G, Liang G, Aparicio A, Jones PA: **Epigenetics in human disease and prospects for epigenetic therapy.** *Nature* 2004, **429**(6990):457-463.
19. Cedoz PL, Prunello M, Brennan K, Gevaert O: **MethylMix 2.0: an R package for identifying DNA methylation genes.** *Bioinformatics* 2018, **34**(17):3044-3046.
20. Gevaert O: **MethylMix: an R package for identifying DNA methylation-driven genes.** *Bioinformatics* 2015, **31**(11):1839-1841.
21. Wang X, Zhang D, Zhang C, Sun Y: **Identification of epigenetic methylation-driven signature and risk loci associated with survival for colon cancer.** *Annals of translational medicine* 2020, **8**(6):324.
22. Dai QX, Liao YH, Deng XH, Xiao XL, Zhang L, Zhou L: **A novel epigenetic signature to predict recurrence-free survival in patients with colon cancer.** *Clinica chimica acta; international journal of clinical chemistry* 2020, **508**:54-60.

23. Cheng S, Jiang Z, Xiao J, Guo H, Wang Z, Wang Y: **The prognostic value of six survival-related genes in bladder cancer.** *Cell death discovery* 2020, **6**:58.
24. Wang Y, Li J, Shao C, Tang X, Du Y, Xu T, Zhao Z, Hu H, Sheng Y, Hu C *et al*: **Systematic profiling of diagnostic and prognostic value of autophagy-related genes for sarcoma patients.** *BMC cancer* 2021, **21**(1):58.
25. Xia WX, Yu Q, Li GH, Liu YW, Xiao FH, Yang LQ, Rahman ZU, Wang HT, Kong QP: **Identification of four hub genes associated with adrenocortical carcinoma progression by WGCNA.** *PeerJ* 2019, **7**:e6555.
26. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic acids research* 2015, **43**(7):e47.
27. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, Teschendorff AE: **ChAMP: updated methylation analysis pipeline for Illumina BeadChips.** *Bioinformatics* 2017, **33**(24):3982-3984.
28. Pan F, Chen T, Sun X, Li K, Jiang X, Forsti A, Zhu Y, Lai M: **Prognosis Prediction of Colorectal Cancer Using Gene Expression Profiles.** *Frontiers in oncology* 2019, **9**:252.
29. Wen S, He L, Zhong Z, Mi H, Liu F: **Prognostic Model of Colorectal Cancer Constructed by Eight Immune-Related Genes.** *Frontiers in molecular biosciences* 2020, **7**:604252.
30. Zhao QQ, Jiang C, Gao Q, Zhang YY, Wang G, Chen XP, Wu SB, Tang J: **Gene expression and methylation profiles identified CXCL3 and CXCL8 as key genes for diagnosis and prognosis of colon adenocarcinoma.** *Journal of cellular physiology* 2020, **235**(5):4902-4912.
31. Huang L, Chen C, Zhang G, Ju Y, Zhang J, Wang H, Li J: **STK33 overexpression in hypopharyngeal squamous cell carcinoma: possible role in tumorigenesis.** *BMC cancer* 2015, **15**:13.
32. Yang T, Song B, Zhang J, Yang GS, Zhang H, Yu WF, Wu MC, Lu JH, Shen F: **STK33 promotes hepatocellular carcinoma through binding to c-Myc.** *Gut* 2016, **65**(1):124-133.
33. Wang P, Cheng H, Wu J, Yan A, Zhang L: **STK33 plays an important positive role in the development of human large cell lung cancers with variable metastatic potential.** *Acta biochimica et biophysica Sinica* 2015, **47**(3):214-223.
34. Yin MD, Ma SP, Liu F, Chen YZ: **Role of serine/threonine kinase 33 methylation in colorectal cancer and its clinical significance.** *Oncology letters* 2018, **15**(2):2153-2160.
35. Flebbe H, Hamdan FH, Kari V, Kitz J, Gaedcke J, Ghadimi BM, Johnsen SA, Grade M: **Epigenome Mapping Identifies Tumor-Specific Gene Expression in Primary Rectal Cancer.** *Cancers (Basel)* 2019, **11**(8).

36. Roberts DL, O'Dwyer ST, Stern PL, Renehan AG: **Global gene expression in pseudomyxoma peritonei, with parallel development of two immortalized cell lines.** *Oncotarget* 2015, **6**(13):10786-10800.
37. Martinez C, Rodino-Janeiro BK, Lobo B, Stanifer ML, Klaus B, Granzow M, Gonzalez-Castro AM, Salvo-Romero E, Alonso-Cotner C, Pigrau M *et al*: **miR-16 and miR-125b are involved in barrier function dysregulation through the modulation of claudin-2 and cingulin expression in the jejunum in IBS with diarrhoea.** *Gut* 2017, **66**(9):1537-1538.
38. Ashikari D, Takayama K, Obinata D, Takahashi S, Inoue S: **CLDN8, an androgen-regulated gene, promotes prostate cancer cell proliferation and migration.** *Cancer Sci* 2017, **108**(7):1386-1393.
39. Wu J, Gao F, Xu T, Li J, Hu Z, Wang C, Long Y, He X, Deng X, Ren D *et al*: **CLDN1 induces autophagy to promote proliferation and metastasis of esophageal squamous carcinoma through AMPK/STAT1/ULK1 signaling.** *Journal of cellular physiology* 2020, **235**(3):2245-2259.
40. Kerachian MA, Javadmanesh A, Azghandi M, Shariatpanahi AM, Yassi M, Davodly ES, Talebi A, Khadangi F, Soltani G, Hayatbakhsh A *et al*: **Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor.** *Scientific reports* 2020, **10**(1).
41. Feodorova Y, Tashkova D, Koev I, Todorov A, Kostov G, Simitchiev K, Belovejdov V, Dimov R, Sarafian V: **Novel insights into transcriptional dysregulation in colorectal cancer.** *Neoplasma* 2018, **65**(3):415-424.
42. Li M, Zhao LM, Li SL, Li J, Gao B, Wang FF, Wang SP, Hu XH, Cao J, Wang GY: **Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients.** *Cancer medicine* 2018, **7**(9):4650-4664.
43. Liu GJ, Wang YJ, Yue M, Zhao LM, Guo YD, Liu YP, Yang HC, Liu F, Zhang X, Zhi LH *et al*: **High expression of TCN1 is a negative prognostic biomarker and can predict neoadjuvant chemosensitivity of colon cancer.** *Scientific reports* 2020, **10**(1).
44. Jiang L, Tolani B, Yeh CC, Fan Y, Reza JA, Horvai AE, Xia E, Kratz JR, Jablons DM, Mann MJ: **Differential gene expression identifies KRT7 and MUC1 as potential metastasis-specific targets in sarcoma.** *Cancer management and research* 2019, **11**:8209-8218.
45. Wang JY, Wang CL, Wang XM, Liu FJ: **Comprehensive analysis of microRNA/mRNA signature in colon adenocarcinoma.** *European review for medical and pharmacological sciences* 2017, **21**(9):2114-2129.
46. Giordano G, Parcesepe P, D'Andrea MR, Coppola L, Di Raimo T, Remo A, Manfrin E, Fiorini C, Scarpa A, Amoreo CA *et al*: **JAK/Stat5-mediated subtype-specific lymphocyte antigen 6 complex, locus G6D (LY6G6D) expression drives mismatch repair proficient colorectal cancer.** *Journal of experimental & clinical cancer research : CR* 2019, **38**(1):28.

47. Guan J, Umapathy G, Yamazaki Y, Wolfstetter G, Mendoza P, Pfeifer K, Mohammed A, Hugosson F, Zhang H, Hsu AW *et al*: **FAM150A and FAM150B are activating ligands for anaplastic lymphoma kinase.** *eLife* 2015, **4**:e09811.

48. Janostiak R, Malvi P, Wajapeyee N: **Anaplastic Lymphoma Kinase Confers Resistance to BRAF Kinase Inhibitors in Melanoma.** *iScience* 2019, **16**:453-467.

49. Hallberg B, Palmer RH: **Mechanistic insight into ALK receptor tyrosine kinase in human cancer biology.** *Nat Rev Cancer* 2013, **13**(10):685-700.

50. Yotani T, Yamada Y, Arai E, Tian Y, Gotoh M, Komiyama M, Fujimoto H, Sakamoto M, Kanai Y: **Novel method for DNA methylation analysis using high-performance liquid chromatography and its clinical application.** *Cancer Sci* 2018, **109**(5):1690-1700.

Tables

Table 1. Clinical information of TCGA and GEO dataset

sample details for DEGs from TCGA			sample details for DMGs from TCGA			GSE39582 for validation		
Tumor	Normal	Overall	Tumor	Normal	Overall	Tumor	Normal	Overall
(N=616)	(N=51)	(N=667)	(N=388)	(N=45)	(N=433)	(N=566)	(N=19)	(N=585)
288 (46.8%)	28 (54.9%)	316 (47.4%)	179 (46.1%)	21 (46.7%)	200 (46.2%)	256 (45.2%)	7 (36.8%)	263 (45.0%)
328 (53.2%)	23 (45.1%)	351 (52.6%)	209 (53.9%)	24 (53.3%)	233 (53.8%)	310 (54.8%)	12 (63.2%)	322 (55.0%)
192 (31.2%)	14 (27.5%)	206 (30.9%)	148 (38.1%)	11 (24.4%)	159 (36.7%)	157 (27.7%)	3 (15.8%)	160 (27.4%)
424 (68.8%)	37 (72.5%)	461 (69.1%)	240 (61.9%)	34 (75.6%)	274 (63.3%)	409 (72.3%)	16 (84.2%)	425 (72.6%)
stage								
103 (16.7%)	8 (15.7%)	111 (16.6%)	53 (13.7%)	5 (11.1%)	58 (13.4%)	33 (5.8%)	5 (26.3%)	38 (6.5%)
228 (37.0%)	24 (47.1%)	252 (37.8%)	144 (37.1%)	21 (46.7%)	165 (38.1%)	264 (46.6%)	7 (36.8%)	271 (46.3%)
178 (28.9%)	9 (17.6%)	187 (28.0%)	119 (30.7%)	10 (22.2%)	129 (29.8%)	205 (36.2%)	5 (26.3%)	210 (35.9%)
86 (14.0%)	9 (17.6%)	95 (14.2%)	52 (13.4%)	9 (20.0%)	61 (14.1%)	60 (10.6%)	0 (0%)	60 (10.3%)
21 (3.4%)	1 (2.0%)	22 (3.3%)	20 (5.2%)	0 (0%)	20 (4.6%)	4 (0.7%)	2 (10.5%)	6 (1.0%)
19 (3.1%)	2 (3.9%)	21 (3.1%)	10 (2.6%)	0 (0%)	10 (2.3%)	11 (1.9%)	1 (5.3%)	12 (2.1%)
104 (16.9%)	7 (13.7%)	111 (16.6%)	54 (13.9%)	5 (11.1%)	59 (13.6%)	45 (8.0%)	4 (21.1%)	49 (8.4%)
421 (68.3%)	36 (70.6%)	457 (68.5%)	271 (69.8%)	37 (82.2%)	308 (71.1%)	367 (64.8%)	12 (63.2%)	379 (64.8%)
69 (11.2%)	6 (11.8%)	75 (11.2%)	50 (12.9%)	3 (6.7%)	53 (12.2%)	119 (21.0%)	0 (0%)	119 (20.3%)
3 (0.5%)	0 (0%)	3 (0.4%)	3 (0.8%)	0 (0%)	3 (0.7%)	24 (4.2%)	2 (10.5%)	26 (4.4%)
455 (73.9%)	35 (68.6%)	490 (73.5%)	263 (67.8%)	27 (60.0%)	290 (67.0%)	482 (85.2%)	17 (89.5%)	499 (85.3%)
85 (13.8%)	9 (17.6%)	94 (14.1%)	51 (13.1%)	9 (20.0%)	60 (13.9%)	61 (10.8%)	0 (0%)	61 (10.4%)
66 (10.7%)	6 (11.8%)	72 (10.8%)	66 (17.0%)	8 (17.8%)	74 (17.1%)	3 (0.5%)	0 (0%)	3 (0.5%)
10 (1.6%)	1 (2.0%)	11 (1.6%)	8 (2.1%)	1 (2.2%)	9 (2.1%)	20 (3.5%)	2 (10.5%)	22 (3.8%)
349 (56.7%)	34 (66.7%)	383 (57.4%)	211 (54.4%)	28 (62.2%)	239 (55.2%)	302 (53.4%)	12 (63.2%)	314 (53.7%)
147 (23.9%)	8 (15.7%)	155 (23.2%)	101 (26.0%)	10 (22.2%)	111 (25.6%)	134 (23.7%)	3 (15.8%)	137 (23.4%)

116	9	125	72	7	79	98	2	100
(18.8%)	(17.6%)	(18.7%)	(18.6%)	(15.6%)	(18.2%)	(17.3%)	(10.5%)	(17.1%)
4 (0.6%)	0 (0%)	4 (0.6%)	4 (1.0%)	0 (0%)	4 (0.9%)	32 (5.7%)	2 (10.5%)	34 (5.8%)

Abbreviation: AJCC, American Joint Committee on Cancer.

Table 2. The information of MDGs

GeneName	Tumor_AVG ¹	Normal_AVG ²	deltaBeta	P.Value	Cor ³	Cor p-value ⁴
ARHGAP20	0.56	0.13	-0.44	6.90E-17	-0.3423	3.02E-12
CLDN1	0.47	0.91	0.44	1.66E-33	-0.4262	8.90E-19
EPHX4	0.61	0.97	0.36	1.08E-19	-0.4539	2.26E-21
FAM150A	0.55	0.25	-0.34	4.76E-16	-0.4362	1.09E-19
KCNJ12	0.33	0.02	-0.31	1.21E-17	-0.4154	8.02E-18
KRT7	0.73	0.31	-0.51	1.59E-38	-0.454	2.20E-21
LY6G6D	0.40	0.76	0.32	4.07E-36	-0.3857	2.17E-15
SPTBN5	0.61	0.91	0.31	1.40E-28	-0.3836	3.16E-15
STK33	0.46	0.11	-0.36	9.67E-17	-0.5458	6.91E-32
TCN1	0.44	0.81	0.36	2.94E-26	-0.4672	1.07E-22

¹Tumor_AVG, average methylation degree of tumor samples.

²Normal_AVG, average methylation degree of normal samples.

³Cor, Correlation between methylation degree and gene expression.

⁴Cor p-value, the test value of correlation.

Table 3. Statistical information of expression and methylation values in MDGs.

GeneName	Expression datasets				
	dataset for discovery	datasets for validation			
	TCGA(READ&COAD)	GSE44076	GSE21815	GSE28000	GSE39582
ARHGAP20	-0.75	-0.76	-1.74	-0.84	-0.90
KCNJ12	-0.61	-0.32	-1.17	-1.04	/
STK33	-0.59	-0.01	-0.44	-1.29	-0.19
KRT7	0.60	1.22	0.58	0.24	0.70
SPTBN5	0.60	0.18	1.67	0.50	0.34
CLDN1	0.77	4.92	5.06	3.14	3.98
EPHX4	1.11	2.80	4.50	1.82	2.45
FAM150A	1.16	1.40	1.63	1.69	/
LY6G6D	1.48	3.84	/	/	/
TCN1	1.56	1.91	3.42	2.20	1.61
GeneName	Methylation datasets				
	dataset for discovery	datasets for validation			
	TCGA	GSE48684	GSE53051	GSE77718	GSE101764
KRT7	-0.51	-0.38	-0.30	-0.34	-0.32
ARHGAP20	-0.44	-0.38	-0.29	-0.38	-0.31
STK33	-0.36	-0.36	-0.25	-0.37	-0.26
FAM150A	-0.34	-0.34	-0.25	-0.34	-0.26
KCNJ12	-0.31	-0.25	-0.23	-0.33	-0.23
SPTBN5	0.31	0.30	0.22	0.21	0.23
LY6G6D	0.32	0.17	0.27	0.19	0.23
TCN1	0.36	0.25	0.29	0.26	0.23
EPHX4	0.36	0.22	0.20	0.14	0.21
CLDN1	0.44	0.22	0.35	0.23	0.31

The number of MDGs in expression datasets means **logFC** value, while in methylation means **deltaBeta** value.

Table 4. The performance of the MDGs as independent indicator using RF, LR and SVM algorithm.

GeneName	RF algorithm		LR algorithm		SVM algorithm	
	AUC	P-value	AUC	P-value	AUC	P-value
CLDN1	0.988	0	0.999	0	0.978	0
EPHX4	0.971	0	0.991	0	0.959	1.55E-246
TCN1	0.966	1.18E-214	0.979	0	0.932	1.56E-129
ARHGAP20	0.921	2.12E-141	0.915	1.03E-88	0.846	9.53E-45
LY6G6D	0.886	2.32E-81	0.772	2.15E-20	0.763	3.35E-24
FAM150A	0.869	1.82E-41	0.885	1.82E-62	0.815	1.24E-32
KCNJ12	0.857	5.60E-50	0.778	1.36E-19	0.732	3.88E-15
KRT7	0.860	4.01E-44	0.858	2.12E-53	0.727	1.26E-14
STK33	0.842	1.78E-39	0.841	1.20E-39	0.721	2.96E-14
SPTBN5	0.678	3.19E-07	0.596	0.007856458	0.5	1

Abbreviation:

RF: Random Forest; LR: logistic regression; SVM: Support vector machine.

Table 5. The importance of diagnostic biomarkers

Predict factor	Normal	Tumor	MeanDecreaseAccuracy	MeanDecreaseGini
CLDN1	0.29	0.10	0.17	115.72
EPHX4	0.17	0.04	0.09	87.21
TCN1	0.08	0.01	0.04	49.64
ARHGAP20	0.12	0.03	0.07	41.59
FAM150A	0.01	0.01	0.01	14.79
LY6G6D	0.05	0.00	0.02	10.32
KRT7	0.02	0.00	0.01	6.76
STK33	0.04	0.01	0.02	6.65
KCNJ12	0.06	0.00	0.02	6.58
SPTBN5	0.00	0.00	0.00	0.72

The **MeanDecreaseAccuracy** and **MeanDecreaseGini** of each indicator was produced by the importance function in Random Forest model.

Figures

Figure 1

Workflow chart of data generation and analysis.

Figure 2

Identification of candidate biomarkers

- (A) Volcano plot on DEGs. The green dots represent downregulated genes, while the red dots represent upregulated genes.
- (B) Statistical differences of the DMGs between normal and tumor groups were compared through the unpaired t test. The values of P are labeled above each boxplot with asterisks. (*P < 0.05, ** P < 0.01, *** P < 0.001, **** P < 0.0001).
- (C) The correlations between each candidate biomarker. The color of the number in the grid represents the strength of correlation between the corresponding genes.

Unsupervised hierarchical clustering of the MDGs for use in the diagnosis prediction model in the training (D) and validation datasets (E) from TCGA and GTEx.

(F, G) Unsupervised hierarchical clustering of seven DEGs for use in the diagnosis prediction model in the training dataset from TCGA and GTEx (F) and validation dataset from GEO (G).

Figure 3

Correlation between the expression and methylation degree of the ten MDGs. X-axis represents the degree of methylation and Y-axis represents the gene expression level.

Figure 4

ROC of the diagnostic prediction model with candidate biomarkers using RF algorithm.

(A-J) independent signature.

(K, L) combined DMGs panel. (K) ten MDGs, (L) seven MDGs.

Figure 5

Identification and assessment of prognostic biomarkers

(A) Univariable Cox proportional hazards regression analysis of the relation between the candidate MDGs and OS status.

(B, C) Kaplan-Meier curves indicate that two MDGs significantly related OS (B) STK33, (C) EPHX4

(D) Kaplan-Meier curve shows that OS in the low score group was significantly higher than that in the high score group.

(E) Time-dependent ROC curve analysis of the prognostic biomarkers.

Figure 6

Construction and validation of a nomogram. (A) Nomogram to predict the probability of OS in 1, 3 and 5 years for CRC. (B) Forest plots showing the associations between patient's characteristics and OS. (C-E) Calibration plot of the nomogram to predict the probability of OS at 1, 3 and 5 years.

Figure 7

Figure legend not provided with this version.

Supplementary Files

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

- [TableS1.xls](#)
- [TableS2.xls](#)
- [FigureS1.png](#)
- [FigureS2.png](#)
- [FigureS3.png](#)