



OPEN

Analyzing aberrant DNA methylation in colorectal cancer uncovered intangible heterogeneity of gene effects in the survival time of patients

Saeedeh Hajebi Khaniki^{1,2}, Farhad Shokoohi^{2✉}, Habibollah Esmaily^{1,3} & Mohammad Amin Kerachian⁴

Colorectal cancer (CRC) involves epigenetic alterations. Irregular gene-methylation alteration causes and advances CRC tumor growth. Detecting differentially methylated genes (DMGs) in CRC and patient survival time paves the way to early cancer detection and prognosis. However, CRC data including survival times are heterogeneous. Almost all studies tend to ignore the heterogeneity of DMG effects on survival. To this end, we utilized a sparse estimation method in the finite mixture of accelerated failure time (AFT) regression models to capture such heterogeneity. We analyzed a dataset of CRC and normal colon tissues and identified 3406 DMGs. Analysis of overlapped DMGs with several Gene Expression Omnibus datasets led to 917 hypo- and 654 hyper-methylated DMGs. CRC pathways were revealed via gene ontology enrichment. Hub genes were selected based on Protein–Protein-Interaction network including *SEMA7A*, *GATA4*, *LHX2*, *SOST*, and *CTLA4*, regulating the Wnt signaling pathway. The relationship between identified DMGs/hub genes and patient survival time uncovered a two-component mixture of AFT regression model. The genes *NMNAT2*, *ZFP42*, *NPAS2*, *MYLK3*, *NUDT13*, *KIRREL3*, and *FKBP6* and hub genes *SOST*, *NFATC1*, and *TLE4* were associated with survival time in the most aggressive form of the disease that can serve as potential diagnostic targets for early CRC detection.

Colorectal cancer (CRC), the third most common cancer worldwide, is a group of diseases characterized by genetic and epigenetic changes^{1,2}. Despite being the second leading cause of cancer-related deaths, less attention has been paid to early detection due to the fact that patients do not adhere to invasive screening tests such as colonoscopy³. It has been shown that epigenetic alterations in solid and liquid biopsies can be used for early detection and thus prognosis and effective treatment⁴. DNA methylation at CpG sites (5mc) is an epigenetic mark that regulates gene expression through transcriptional silencing⁵. Aberrant DNA methylation plays a crucial role in the pathogenesis and progression of CRC and has emerged as a promising diagnostic marker for the disease⁶. In particular, aberrant DNA methylation can impact genes where their inactivation may exacerbate tumor formation through the induction of genomic instability or by directly silencing the methylated gene⁷.

Much research has been done to develop comprehensive panels of biomarkers based on DNA methylation that can facilitate accurate diagnosis of CRC⁸. While the genes *SEPT9*, *NDRG4*, and *BMP3* are FDA-approved for CRC^{9,10}, there are many other genes such as *APC*, *SFRP1*, *TFPI2*, and *VIM* that have not yet been approved⁸.

In order to detect and validate genes that are potential CRC biomarkers, the following steps should be taken. Firstly, a panel of biomarkers must be developed using accurate statistical methods with a deep understanding of the underlying biology of the disease and the molecular mechanisms that drive them. Secondly, the significant biomarkers must be validated via in silico validation using several other datasets; and thirdly, the effectiveness of top candidate biomarkers in improving patient health should be verified using survival models. Lack of adequate

¹Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran. ²Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. ³Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. ⁴Medical Genetics Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. ✉email: farhad.shokoohi@unlv.edu

precision in each of the above steps leads to misleading conclusions. Among others, two issues affect precision: removing genomic positions with missing values or low read-depth and ignoring the heterogeneity of DMG effects on survival times.

To accurately predict the differentially methylated profiles in CRC, one must consider all biological and environmental factors such as dietary¹¹, aging¹², and hazardous behaviors¹³ (e.g., smoking), among others. Such factors are often ignored by most studies when predicting methylation profiles. In addition, methylation data always suffer from heavy missing values that can affect subsequent analyses. For instance, 68% of CpG sites have missing values in at least one sample in our dataset (Section 2). Almost all DNA methylation pipelines, except a few such as the DMCHMM method¹⁴, filter out such positions from the analysis. We used DMCHMM to not only account for extra covariates but also efficiently impute the missing values.

Having identified the differentially methylated genes (DMG) associated with CRC and validating them, it is crucial to identify their underlying signaling pathways that regulate gene expression^{15,16}. The main known CRC pathways are Wnt¹⁷, MAPK¹⁸, TGF- β ¹⁹, and TP53²⁰. Although significant progress has been made in understanding the biology of CRC, there are still many unknown pathways and mechanisms involved in this disease. Identification of hub genes, also known as driver genes is the next step in the analysis of biomarker detection. Hub genes play a critical role in regulating several genes in the biological network and have the potential to be regarded as therapeutic targets in CRC²¹.

In the next step, the relationship between identified DMGs and the survival time of CRC patients should be evaluated. Most studies employ a limited panel of biomarkers selected through conventional univariate Cox proportional hazard regression models and overlook the potential effects of the rest of the biomarkers^{22–24}. In a recent study²⁵, the Cox-LASSO survival model was used to account for a larger set of biomarkers but ignored the heterogeneity of covariate effects. To the best of our knowledge, none of the studies have taken into account the heterogeneity of DMG effects on survival time. To address this problem, one may use the sparse estimation method in the finite mixture of accelerated failure time (AFT) regression models²⁶. Prior to this step, it is common to screen the number of genes to a manageable magnitude. This process can be done by selecting the top highly correlated genes with survival time of the patients using the correlation-adjusted scoring method²⁷.

This study aimed to identify CRC-related DMGs to serve as potential biomarkers for early detection by including all the available information in the data and avoiding the exclusion of any genomic position. To this end, we acquired a high-throughput DNA methylation dataset which consists of patients with CRC and healthy individuals. Information on age, history of smoking, and drug abuse was also collected. A description of the data is provided in Section “Methods”. Information on other datasets used for validation and survival analysis and all statistical and Bioinformatics methods are listed in this section. In Section “Results”, a comprehensive analysis of data is conducted. Section “Discussion” gives a discussion and some concluding remarks.

Methods

In this section, we outline the data analysis process we followed to detect DMGs, hub genes, and their effects on the survival time and enriched pathways of CRC. Figure 1 depicts the flowchart of this process.

Phase I (pre-processing of discovery samples)

To identify methylation-based CRC biomarkers, information on 6 patients with adenocarcinoma of CRC and 6 normal males was obtained. Two groups were matched based on age, and family history of cancer²⁸.

The methylation profiles in our dataset are derived from a three-step pre-processing phase conducted through SureSelectXT Human Methyl-Seq. Initially, the purity and quantity of 12 DNA tissue samples were assessed using specific criteria, including a minimum concentration of 50ng/ μ l, a purity ratio (A_{260}/A_{280}) ≥ 1.7 , a volume of at least 20ng/ μ l, and a total amount exceeding 3.0 μ g. Subsequently, global methylation profiles of CRC and normal samples were analyzed using SureSelectXT Human Methyl-Seq.

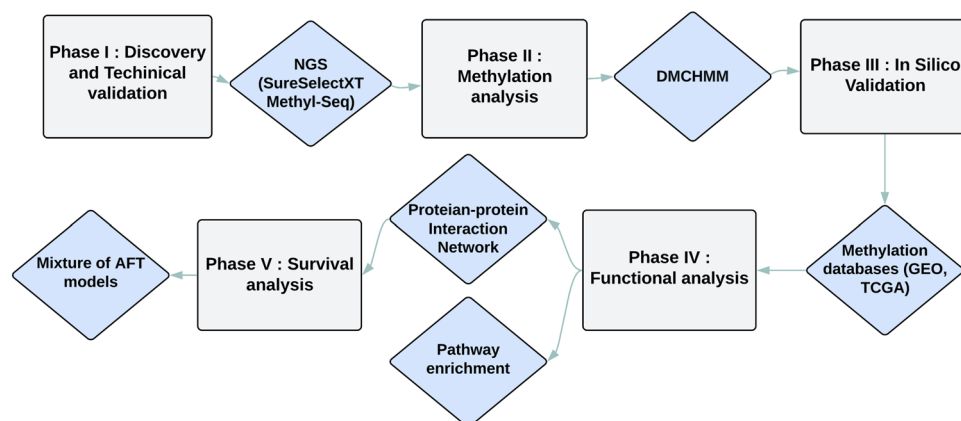


Figure 1. Study workflow for the analysis of CRC datasets.

Pre-sequencing tasks, such as sample collection and DNA extraction, were consistently carried out by a single technician. Experimental conditions for all samples remained constant both before and after sequencing. During the sequencing process, all sample runs were executed simultaneously using the same device, employing Next-generation sequencing technology, a highly parallel sequencing method. This approach minimized the potential introduction of batch effects attributable to non-biological factors such as variations in laboratory conditions, personnel, and equipment used in the experiment.

In the second step, a quality control assessment of total reads using FastQ²⁹ was conducted. This step aimed to provide informative global and graphical representations of read quality in methylation sequencing, both pre and post-alignment. Notably, our data consistently exhibited high quality in raw sequencing reads across all samples. Subsequently, Trim Galore³⁰ was utilized to process the raw sequencing reads. This involved the removal of sequencing adapters, specifically the Illumina universal adapter, and discarding the low-quality bases (those with quality scores below 67, as per Illumina standards) located at the 3' end of reads. Additionally, any ambiguous bases found in both reads were removed.

Finally, the raw bisulfite sequencing data were aligned to the human reference genome (GRCh37/19) using Bismark³¹. Several comparisons and visualization confirmed minimal to no presence of batch effects in our data. This discovery dataset includes methylation read counts and read-depth for each CpG site, generating 57 to 76 million Illumina sequencing reads per subject. Between 88.5% and 89.8% of sequenced reads were mapped to either strand of the human genome (GRCh37/19). On average, each CpG site was sequenced between 19 × and 24 × per sample. The sequencing details for the subjects are presented in Table 1. Approximately 68% of the 19,530,818 CpG sites have missing information in at least one sample.

Phase II (identification of differentially methylated genes)

We utilized the DMCHMM pipeline³² to identify CpGs with differentially methylated patterns between CRC and normal discovery samples. We specifically did not remove any position with missing information or low read-depth. The missing information was imputed using DMCHMM via hidden Markov models¹⁴. Significant differentially methylated cytosines (DMCs) were selected based on the FDR threshold of 0.05. DMCs were aligned to the human reference genome (GRCh37/19) using the UCSC Genome Browser (<https://genome.ucsc.edu>). A gene whose promoter was mainly hypo- or hyper-methylated was classified as hypo- or hyper DMG, respectively.

Phase III (cross-platform validation)

To validate our result, several methylation profiles (GSE53051³³, GSE77718³⁴, GSE101764¹³, GSE42752³⁵, GSE48684³⁶) were extracted from the Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>). Of these datasets, a total of 212 CRC and 242 normal mucosa tissue samples were selected based on setup conditions to minimize the confounding effect of other variables. These datasets have provided valuable insights into the molecular alterations that occur in CRC, and their findings have implications for the diagnosis and treatment of this disease. For the analysis of methyl array profiles of validation sets, the GEO2R (<http://www.ncbi.nlm.nih.gov/geo/geo2r/>) web tool and the limma R-package³⁷ were used. To mitigate batch effects, we applied the 'removeBatchEffect' option from the package. A probe was considered differentially methylated if its adjusted p-value was less than 0.05, and the absolute of log₂ of methylation fold change was greater or equal to 1. The differentially methylated probes were aligned to the human reference genome (GRCh37/19) using the FDb.InfiniumMethylation.hg19 package³⁸. In the last step, we compared the lists of DMGs based on the validation sets and our discovery samples to identify consistent hypo/hyper-methylated genes across different populations and platforms.

Phase IV (network construction and functional analysis)

In order to investigate the Protein-Protein Interaction (PPI) network and module analysis, we utilized the 'Search Tool for the Retrieval of Interacting Genes' (STRING) database. We set the interaction score threshold to 0.4 to screen for high-confidence interactions and visualized the resulting network using the Cytoscape³⁹ software

Sample	Total reads	Mapping rate	Methylation (%)	Average coverage	GC (%)
T65	76,723,684	88.50	47.70	24.15	27.04
N16	70,443,130	88.70	45.70	23.53	27.26
T20	67,394,464	88.90	44.70	19.58	27.03
N4	68,165,382	88.80	46.50	22.19	27.19
T31	61,789,306	89.00	46.90	21.69	26.92
N10	57,311,634	89.05	46.70	19.26	27.04
T35	79,004,644	88.90	46.10	24.43	27.11
N7	75,663,274	89.00	47.20	22.62	27.04
T45	64,188,480	89.00	47.40	21.22	27.06
N8	57,091,968	89.80	46.80	20.42	27.41
T67	61,203,576	89.30	44.30	20.77	27.17
N14	66,871,860	89.60	47.40	22.17	27.11

Table 1. Summary statistics of methylation sequencing reads of discovery samples.

(Version 3.9.1). Next, we employed the Molecular Complex Detection (MCODE) algorithm to uncover densely connected substructures within the network. The MCODE score must be greater than 3 and the minimum number of nodes must be 4. In order to identify key hub genes within the network, we used the `cytoHubba` plugin and considered the degree of centrality as a parameter.

To gain insight into the biological mechanisms that are driving CRC and prioritize identified DMGs, we performed functional and pathway enrichment analysis using DAVID⁴⁰ (<https://david.ncifcrf.gov/>). Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG⁴¹) pathways were considered significantly enriched if the p-values were less than 0.05 and the q-values were less than 0.1. The visualization of the identified GO terms and KEGG pathways were done with the `clusterProfiler`⁴², `pathfindR`⁴³, and `ShinyGO`^{44,45} (<http://bioinformatics.sdstate.edu/go/>) packages.

Phase V (Uncovering intangible heterogeneity of DMG effects on survival time)

To explore the relationship between identified DMGs and survival time, the DNA methylation profiles of 521 samples were obtained from The Cancer Genome Atlas (TCGA) network⁴⁶. Complete information on clinical variables including days to follow-up and the status of the patient were analyzed.

We conducted several preliminary analyses on the overall survival time of patients with CRC. First, we estimated the density of the logarithm of the survival times using the Kaplan-Meier estimator. The density plot in Figure 2 shows a mixture distribution. Second, we applied mixture and non-mixture models of normal distributions using the `mixtools` package⁴⁷. The BICs for the mixture of components $K = 1, 2, 3, 4,$ and 5 were estimated as 777.73, 709.02, 712.31, 722.51, and 721.68, respectively, with the lowest BIC observed for the mixture with 2 components. Finally, we employed mixture and non-mixture models of semiparametric scaled data using the stochastic EM algorithm⁴⁸ via `mixtools`. Please note that the package only outputs a mixture with two components, and the BICs for the non-mixture $K = 1$ and the mixture of 2 components were 1323.62 and 582.01, respectively. Similarly, the lowest BIC was observed for the mixture with 2 components. All of these preliminary analyses indicate heterogeneity in the overall survival time of patients with CRC.

Furthermore, a comprehensive literature review revealed numerous conflicting results. For example, researchers⁴⁹ found a significant association between the methylation level of RASSF1A and the overall survival of CRC patients, while other studies⁵⁰ did not observe such an association.

Moreover, several studies indicate stage-specific⁵¹ and age-specific⁵² effects of DNA methylation in certain genes on the survival outcomes of CRC patients. These results clearly suggest heterogeneity in the overall survival time of patients with CRC.

We hypothesize that the effect of identified DMGs and hub genes on the overall survival time varies in each subpopulation, but not all DMGs and hub genes have an effect in each subpopulation, implying that the underlying regression model is sparse. To capture such heterogeneity, we employed the sparse estimation method in the finite mixture of AFT regression models²⁶. The details of the method are given in Section “[Sparse finite mixture of AFT regression models to estimate the DMG effects on the survival times](#)” below. The response variable is “Overall Survival Time”, and the independent variables are the log-transformed average methylation of identified DMGs or hub genes that we discovered through Phase II–IV. The goal of such a regression model is to estimate the effects of each gene in different sub-populations of the response variable, providing insights into the effects of each gene on the survival time of patients with CRC. It is important to note that the response variable (survival time) is subject to right-censoring. The sparse estimation method requires tuning parameters, which are

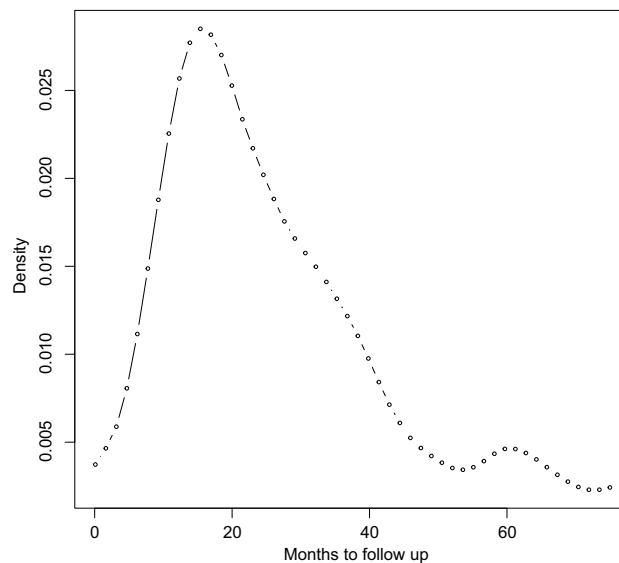


Figure 2. Density estimation of overall survival time (in months) in CRC patients (R package `survPresmooth`, v1.1-11).

estimated using a data-adaptive approach explained in Section “Sparse finite mixture of AFT regression models to estimate the DMG effects on the survival times”.

Sparse finite mixture of AFT regression models to estimate the DMG effects on the survival times

As hypothesized above, the overall survival time of patients with CRC is heterogeneous; thus, we hypothesize that the relationship between overall survival time and DMGs and hub genes found in Phase II-IV is heterogeneous. Such heterogeneity cannot be detected using a regular AFT regression model for censored data. Therefore, we employ the finite mixture of the AFT regression model to capture intangible DMG and hub gene effects on survival time. To this end, we use the finite mixture of AFT regression model:

$$f(y; \theta) \propto \sum_{k=1}^K \pi_k [f_k(y; \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2)]^\delta [S_k(y; \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2)]^{1-\delta},$$

where f_k and S_k are respectively the density of normal distribution and its survival function, $y = \log(t)$, t is the overall survival time, δ is an indicator representing right-censored (i.e., $\delta = 0$ if time is censored and 1 if it is not censored), \mathbf{X} is the vector of all DMGs and hub genes discovered in Phase II-IV, $\boldsymbol{\beta}_k$ is the vector of effects of these genes in Component k of the mixture model, σ_k^2 is the variance, and π_k is the proportion of the k th component.

It is common to screen the number of genes prior to analysis in case of a large number of identified genes. To this end, we applied a correlation-adjusted score method using the `carSurv` package²⁷ to screen the genes.

Next, we used the `fmrs` package⁵³ to fit finite mixture and non-mixture of AFT regression models to the data. We employed the smoothly clipped absolute deviations (SCAD) penalty²⁶. This sparse method requires K tuning parameters which are estimated via the data-adaptive component-wise BIC method proposed in Shokoohi et al.²⁶.

Results

Differentially methylated cytosine detection

We identified 2,691,019 DMCs between CRC and normal groups of the discovery dataset while adjusting for the potential confounding effect of smoking history or drug abuse. Of these identified DMCs, 1,985,557 positions were hypo-methylated and 705,462 CpGs were hyper-methylated in CRC vs normal samples. The heatmaps (see R package `heatmap54`) in Fig. 3a indicate a clear clustering pattern between the CRC and normal samples based on the predicted methylation levels of DMCs.

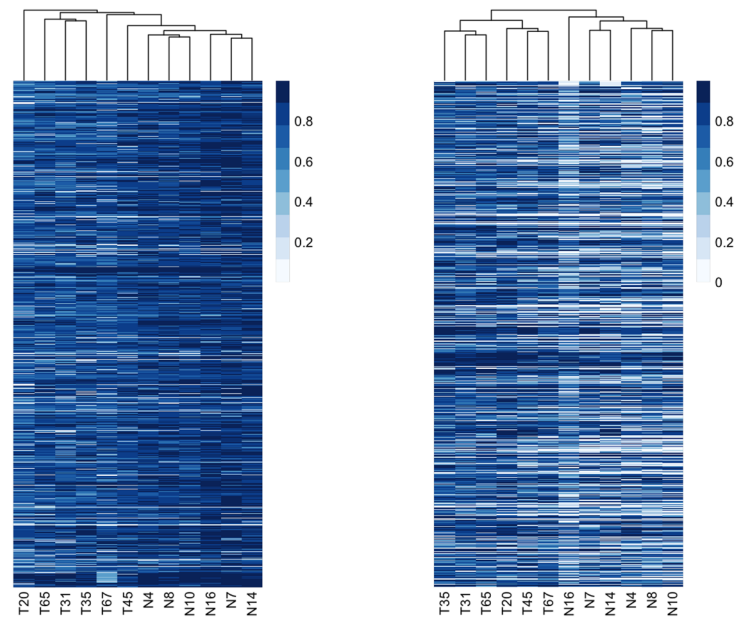
To explore the genomic location of the DMCs, we analyzed their distribution across different regions and summarized the results in Fig. 3b. Intergenic regions were found to harbor the majority of the detected DMCs both in the hypo and hyper categories. Notably, we observed that 32% of hyper-methylated DMCs were located in CpG islands, while only 9% of hypo-methylated DMCs were found in these regions. Additionally, the regions with the highest percentage of hyper-methylated DMCs were identified in introns, exons, and CGI shores. The Chord diagrams (see R package `circulize55`) in Fig. 3c gives a comprehensive overview of how hyper and hypo-methylated DMCs were distributed across different genomic regions. Our findings suggest that many DMCs in intergenic regions were expanded to intronic regions in both hypo and hyper-methylated categories.

Given the potential significance of promoter methylation in cancer development and progression, we focused our subsequent analysis on DMCs located on gene promoters, which encompassed 268,978 CpGs. These CpGs resided on 3406 gene promoters, of which 1394 were hyper-methylated and 2012 were hypo-methylated. The list of DMGs is available as supplementary material.

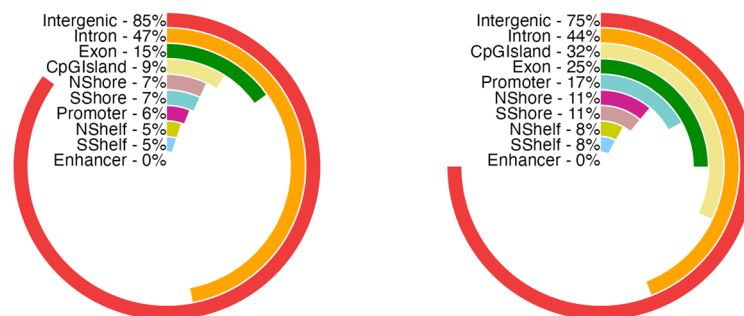
Robust DMGs in CRC

To verify the robustness of identified DMGs, we performed a cross-platform procedure with DMGs identified in selected GEO datasets as depicted in Fig. 4a (see R package `venn56`). The comparison revealed a total of 1571 overlapped DMGs that were consistently identified across multiple studies. As Fig. 4b (see R package `karyoploteR57`) illustrated, the identified DMGs were spread almost evenly across different chromosomes, with chromosomes 1 and 7 having some dense regions of CRC-related DMGs. Within this set, 917 genes were hypo-methylated, and 654 genes were hyper-methylated. We focused our subsequent analysis on these identified DMGs to gain a deeper understanding of their role in CRC pathogenesis.

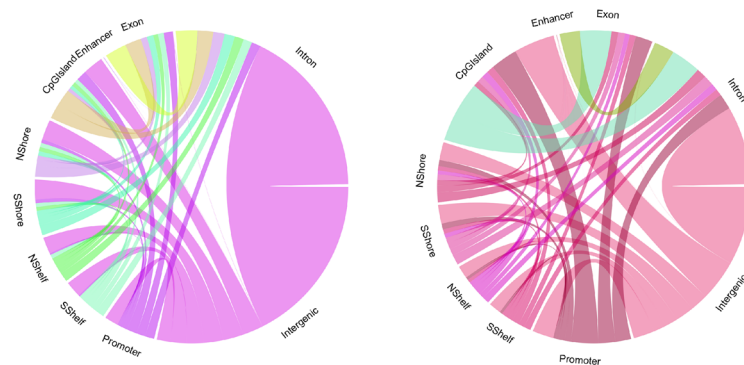
GO enrichment KEGG pathway analysis. The analysis of robust DMGs in CRC utilizing the DAVID tool yielded a variety of enriched biological processes, molecular functions, and cellular components. Specifically, the hyper-methylated DMGs were found to be principally involved in ‘cell fate commitment’, ‘regionalization’, ‘embryonic organ morphogenesis’, ‘embryonic organ development’, ‘pattern specification process’, ‘animal organ morphogenesis’, ‘tube morphogenesis’, ‘tube development’, and ‘neurogenesis’ in the context of biological processes (Fig. 5a, see R Shiny package `ShinyGO44`). Enriched cellular components included ‘basement membrane’, ‘integral component of postsynaptic membrane’, and ‘Collagen-containing extracellular matrix’ (Fig. 5c). Additionally, KEGG pathway analysis indicated that hyper-methylated DMGs were significantly enriched in several pathways, including ‘signaling pathways regulating pluripotency of stem cells’, ‘axon guidance’, ‘morphine addiction’, ‘rap1 signaling pathway’, ‘circadian entrainment’, and ‘pathways in cancer’ (Fig. 5e and Table 2). Regarding biological processes, the hypo-methylated DMGs were found to be associated with a number of processes including ‘keratinization’, ‘keratinocyte differentiation’, ‘epidermal cell differentiation’, and ‘epithelial cell differentiation’ (Fig. 5b). Furthermore, analysis of the cellular component pathway revealed that the hypo-methylated DMGs were most significantly enriched in the ‘cornified envelope’, ‘integral component of the synaptic membrane’, and ‘integral component of the postsynaptic membrane’. Notably, these cellular components demonstrated the highest FDR and fold enrichment (Fig. 5d). Regarding molecular functions, the pathways



(a) Predicted methylation levels of hypo-(left) and hyper-(right) methylated CpGs (R package [pheatmap, v1.0.12](#)).



(b) Genomic location of hypo-(left) and hyper-(right) methylated CpGs (R package [ggplot2, v3.4.4](#)).



(c) Chord diagrams of relations between genomic locations of hypo-(left) and hyper-(right) methylated DMCs (R package [circlize, v0.4.15](#)).

Figure 3. Genomic location of identified differentially methylated CpGs and their predicted levels in CRC (T) and normal (N) samples using DMCHMM. The hierarchical clustering of CRC and normal samples in the heatmaps is based on complete linkage.

with higher fold enrichment included ‘molecular transducer activity’, ‘signaling receptor activity’, and ‘transmembrane signaling receptor activity’. Notably, KEGG pathway analysis revealed that hypo-methylated DMGs were significantly enriched in several pathways, including the ‘oxytocin signaling pathway’, ‘glioma’, ‘adrenergic

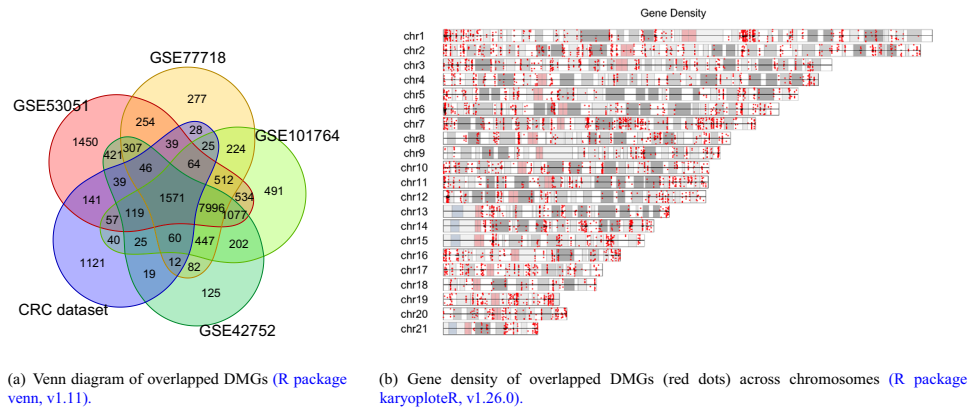


Figure 4. Summary of common identified DMG and their distribution.

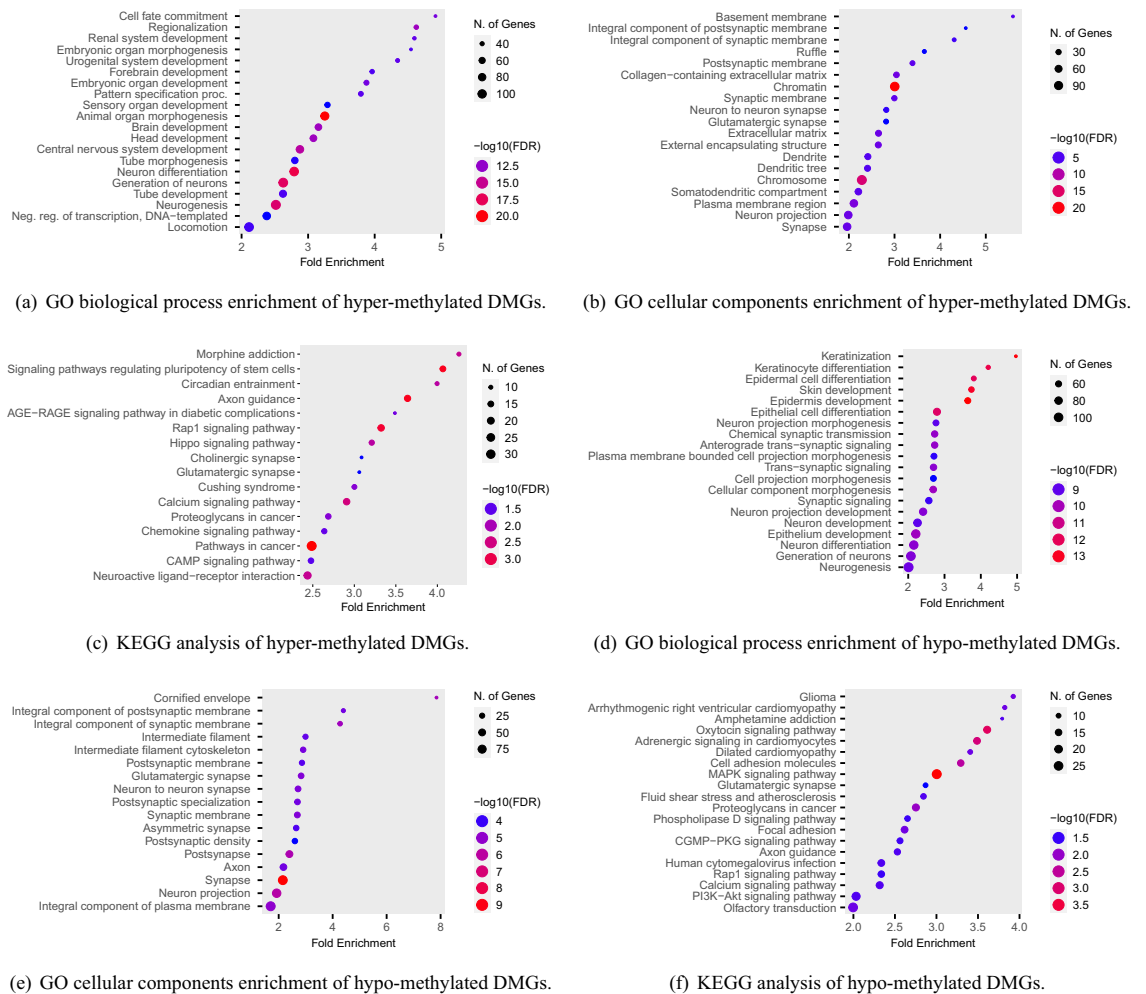


Figure 5. Enrichment analysis of commonly identified DMGs (R Shiny package *ShinyGO*, v0.77).

signaling in cardiomyocytes, ‘MAPK signaling pathway’, ‘arrhythmogenic right ventricular cardiomyopathy’, and ‘cell adhesion molecules’ (Fig. 5f). These results offer valuable insights into the potential mechanisms of DMGs in CRC and identify possible therapeutic targets for this disease. A comprehensive summary of the KEGG pathways of hyper-methylated DMGs can be found in Table 2.

PPI network construction. We ran a PPI network to further investigate the complex interactions between DMGs and find important hub proteins. A total of 606 PPI nodes of the hyper-methylated DMGs were

Enrichment		Pathway		Fold	
FDR	nGenes	Genes	Enrichment	Pathway	Matching proteins in network (labels)
0.0050	10	91	4.26	Morphine addiction	PDE8A, GNAS, SLC32A1, GABRA4, GNGT1, KCNJ3, ADORA1, ADCY1, PRKCB, GNG2
0.0004	15	143	4.07	Signaling pathways regulating pluripotency of stem cells	PAX6, FGFR1, LHX5, HOXA1, MYF5, WNT5A, ID2, BMP4, IGF1R, WNT3A, FZD1, FZD6, AXIN2, ONECUT1, SMAD2
0.0060	10	97	3.99	Circadian entrainment	GNAS, GNGT1, MTNR1B, ITPR1, KCNJ3, ADCY1, PRKCB, GRIN2A, PRKG1, GNG2
0.0004	17	181	3.64	Axon guidance	NEO1, PRKCZ, SEMA5B, NFATC2, CXCL12, UNC5A, WNT5A, EPHA4, SMO, EPHA7, SEMA4F, SEMA6D, SLIT2, ROBO3, UNC5C, SEMA4A, PLXNA4
0.0250	9	100	3.49	AGE-RAGE signaling pathway in diabetic complications	PRKCZ, STAT1, COL4A2, PLCD3, PRKCB, COL4A3, SMAD2, THBD, COL4A1
0.0006	18	210	3.32	Rap1 signaling pathway	PRKCZ, RASGRP2, APBB1IP, FGFR1, GNAS, FGF9, CNR1, VAV3, FGF5, IGF1R, ANGPT1, TIAM1, VAV2, ADCY1, PRKCB, ADORA2B, GRIN2A, SIPA1L1
0.0060	13	157	3.21	Hippo signaling pathway	CTNNA2, PRKCZ, FBXW11, TP73, WNT5A, ID2, BMP4, BMP6, WNT3A, FZD1, FZD6, AXIN2, SMAD2
0.0460	9	113	3.09	Cholinergic synapse	GNGT1, PIK3R5, ITPR1, KCNJ3, ADCY1, PRKCB, CHRM4, CHRM2, GNG2
0.0460	9	114	3.07	Glutamatergic synapse	GNAS, GNGT1, ITPR1, KCNJ3, ADCY1, PRKCB, GRIN2A, GNG2, GRM3
0.0180	12	155	3.00	Cushing syndrome	PDE8A, KCNK2, GNAS, CDK6, CRHR2, WNT5A, ITPR1, WNT3A, FZD1, ADCY1, FZD6, AXIN2
0.0030	18	240	2.91	Calcium signaling pathway	FGFR1, GNAS, FGF9, P2RX3, TACR1, FGF5, GNAL, ITPR1, PLCD3, ADCY1, PRKCB, GDNF, ADORA2B, OXTR, CHRM2, GRIN2A, ATP2A1, HRH1
0.0200	14	202	2.64	Chemokine signaling pathway	PRKCZ, RASGRP2, CXCL12, STAT1, PREX1, GNGT1, VAV3, PIK3R5, TIAM1, VAV2, ADCY1, PRKCB, GNG2
0.0500	11	166	2.56	Wnt signaling pathway	FBXW11, NFATC2, SFRP1, WNT5A, SFRP5, WNT3A, FZD1, SOX17, FZD6, PRKCB, AXIN2
0.0003	34	530	2.49	Pathways in cancer	CTNNA2, RASGRP2, FGFR1, GNAS, MSH2, FGF9, IL7, CDK6, CXCL12, WNT5A, STAT1, BMP4, GNGT1, SMO, RARA, CCNA1, COL4A2, LAMC1, FGF5, IGF1R, PMAIP1, WNT3A, FZD1, ADCY1, FZD6, PRKCB, AXIN2, COL4A3, SMAD2, GNG2, MTF, COL4A1, TXNRD1, NCOA4

Table 2. KEGG pathway analysis of commonly identified hyper-methylated DMGs.

constructed on the basis of the STRING database (Fig. 6, see R Shiny package *ShinyGo*⁴⁴). The 16 node proteins, including *KIT*, *SEMA7A*, *BDNF*, *MEF2A*, *LDB2*, *GATA4*, *LHX2*, *SOST*, *CTLA4*, *NKX2-2*, *TLE4*, *BMP5*, *NFATC1*, *ZFPM1*, *DPYSL2*, and *ITGA2B* that showed a close interaction with other node proteins were chosen as hub genes (Fig. 7a, see *Cytoscape*³⁹). The most important biological process and KEGG pathways of hub genes are shown in Fig. 7b and c. One important module was selected when the number of nodes is greater than 4. The key module demonstrated functions enriched in pathways such as Wnt signaling⁵⁸ (Table 2 and Fig. 8, see R Shiny package *ShinyGo*⁴⁴).

We performed a survival analysis using the TCGA-selected samples to investigate the association of selected hub genes with the survival time of CRC patients. Based on Fig. 9a–d (see GEPIA2021⁵⁹), those patients with gene *SEMA7A* ($p = 0.024$), *SOST* ($p = 0.027$), *NFATC1* ($p = 0.017$), and *TLE4* ($p = 0.0061$) being upregulated, had a significantly lower probability of survival. However, this conclusion is based on univariate analysis, and the effect of other genes and the potential heterogeneity of DMG effects were ignored. We reanalyzed these data by accounting for the heterogeneity of DMG effects and obtained different results as follows.

Intangible heterogeneity of DMG effects on survival time. We studied the relationship between the average promoter methylation of the identified DMGs and the survival time subject to right-censoring by accounting for the heterogeneity of gene effects using an independent set of 521 TCGA CRC samples. To this end, we screened all the 1571 candidate DMGs using the correlation-adjusted regression survival scores to obtain the list of top candidate covariates. This process led to the selection of 95 highly correlated DMGs. These genes were also dysregulated in the TCGA samples. In addition, 4 hub genes that were related to the survival time of CRC patients were added to the list of covariates.

Our analysis yielded a two-component mixture of AFT regression model. The estimated gene effects on the survival time are given in Table 3. The result showed that 46% of the subjects were classified into Component 1, which is the most aggressive form of the disease. Figure 10 (see R package *fmr*⁶⁰) depicts the posterior probability of a subject belonging to Component 1. From this figure, we noticed that all living patients were classified into Component 2, which is the less aggressive form of the disease. A total of 83 and 18 DMGs were active in Components 1 and 2, respectively. Twelve genes including *HLA-F*, *MMP2*, *MT1A*, *RFPL4B*, *SIX6*, *ZFAT*, *BCKDK*, *AMOTL1*, *ADCY10*, *KCNK10*, *STAU2*, and *NOC4L* were not related to survival time in either of the components. These findings demonstrate the heterogeneity of DMG effects in CRC data and justify using a sparse mixture modeling rather than a univariate one. In addition, the DMGs with active promoters in Component 1 can be

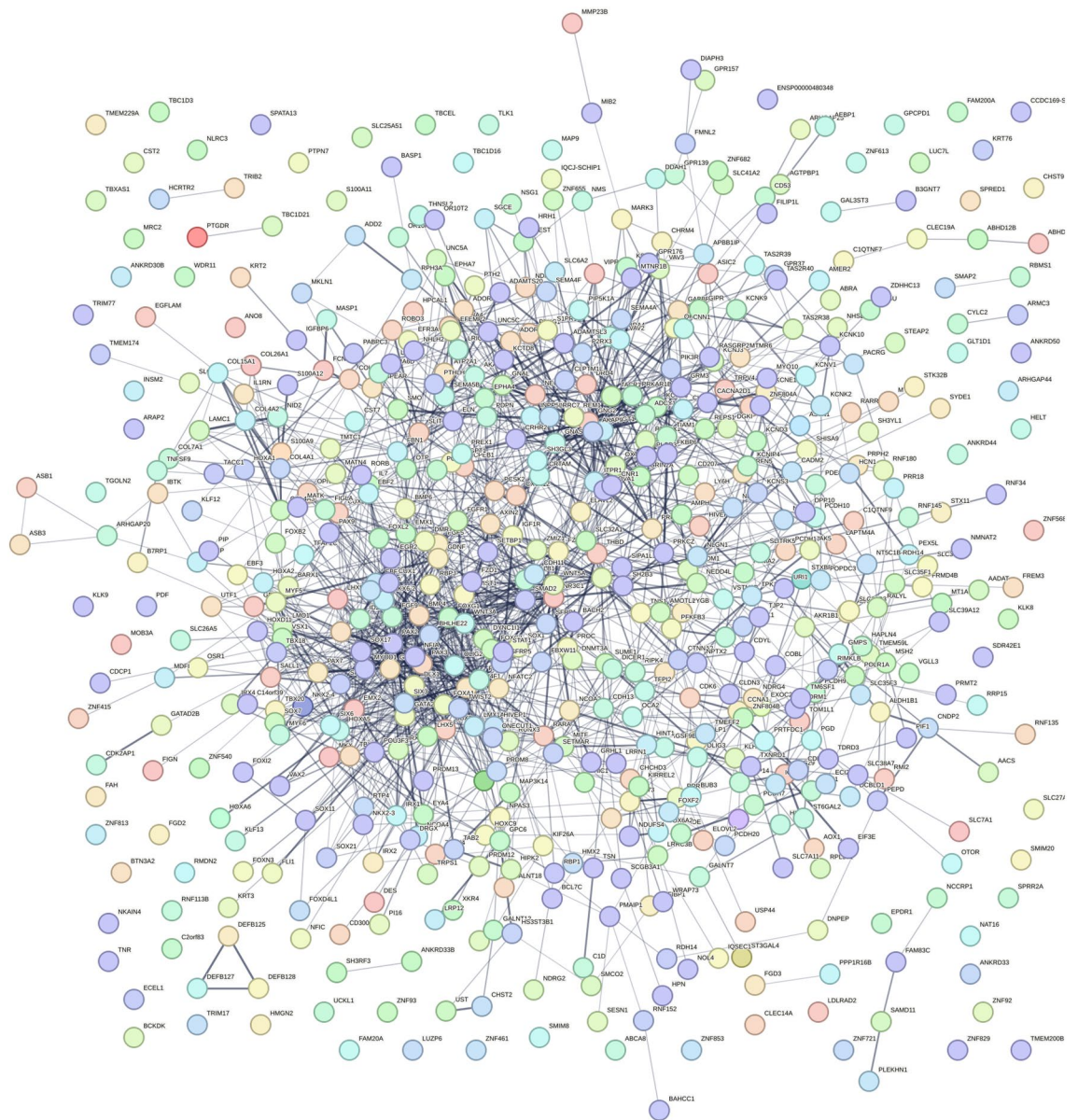


Figure 6. Protein–protein interaction network of hyper-methylated genes. Spots represent the proteins and lines show interactions (R Shiny package *ShinyGO*, v0.77).

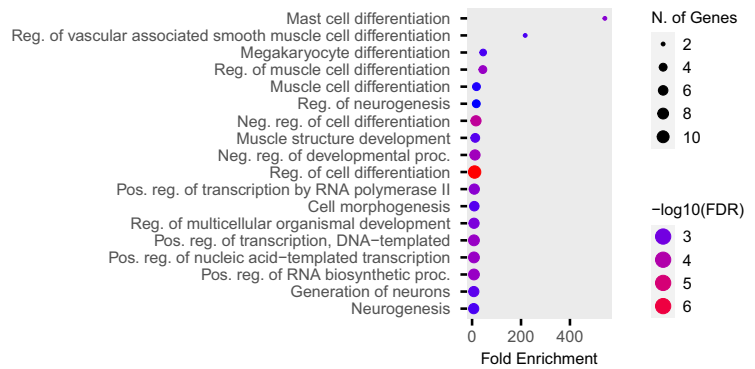
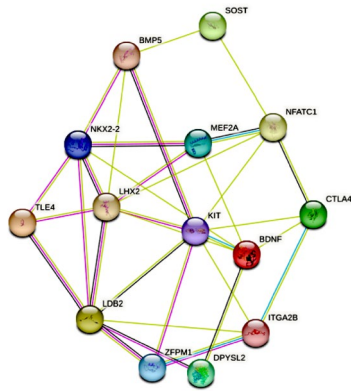
considered biomarkers for CRC prognosis. The bioinformatics and biological information of selected DMGs are given in Table 4.

Discussion

Colorectal cancer is one of the deadliest cancers in the world. Given that early stages of CRC do not display symptoms, proactive screening is the only viable approach to identify the disease⁸³. As DNA methylation changes are closely associated with cancer, their role in CRC biomarker detection in the early stages of cancer is of great importance. Although many CRC biomarkers have been detected in the literature, only a few are used in practice. Our findings resulted in identifying new biomarkers for CRC which can be used for diagnosis and prognosis.

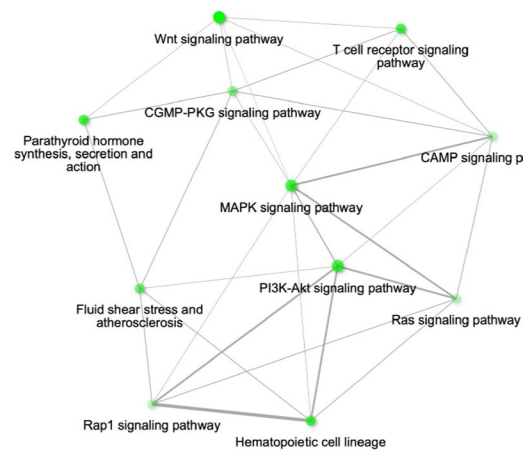
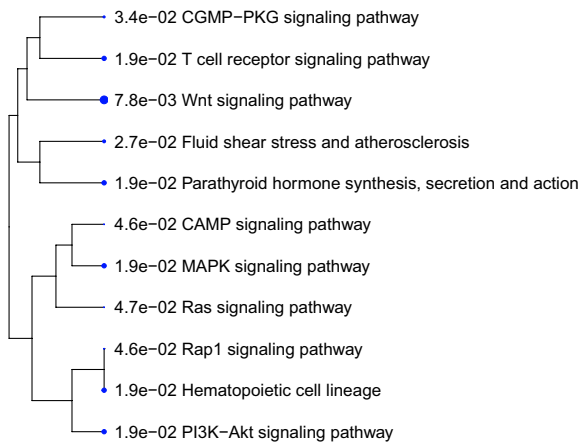
We identified 1,571 DMGs most of which have been previously studied in the literature. Among them, *SEPT9*, *NDRG4*, *VIM*, *APC*, *SFRP1*, *SFRP4*, and *SFRP5*⁸⁴ are the most important CRC-related ones. We also explored CRC-related hub genes. Fourteen functional modules that may play important roles in the early detection of CRC were highlighted and the sub-network of hub genes *KIT*, *SEMA7A*, *BDNF*, *MEF2A*, *LDB2*, *GATA4*, *LHX2*, *SOST*, *CTLA4*, *NKX2-2*, *TLE4*, *BMP5*, *NFATC1*, *ZFPM1*, *DPYSL2*, and *ITGA2B* was extracted. These hub genes were flagged as potential diagnostic and therapeutic targets for CRC in our analysis.

In addition to the diagnostic role of our identified hub genes such as *NKX2-2*, *KIT*, *BDNF*, and *TLE4* in CRC and its sub-types^{74,85–87}, their roles in increasing CRC risk, tumor progression, and targeted therapy have been investigated. For instance, *MEF2A*⁸⁸ and *BMP5*⁸⁹ increase the CRC risk. Up-regulation of the expression



(a) Top module of hyper-methylated hub genes verified by MCODE (<https://cytoscape.org>).

(b) GO biological process enrichment of hyper-methylated hub genes (R Shiny package ShinyGO, v0.77).



(c) Clustered and interaction network of enriched KEGG pathways of hyper-methylated hub genes (R Shiny package ShinyGO, v0.77).

Figure 7. Bioinformatic analysis of hyper-methylated hub genes.

of *ITGB7* and *ITGA2B* has been found to be significantly associated with death by sodium butyrate-induced CRC organoids⁹⁰. Moreover, some studies^{91,92} have shown effective treatments by targeting *CLT-4* and *LDB2n*.

There is a rich literature on the contribution of some of our identified hub genes in CRC and less evidence in support of some others such as *LHX2*, *ZFPM1*, and *DPYSL2*. For instance, the differences in tumor and corresponding adjacent benign tissues regarding *LHX* gene expressions have been investigated⁹³. However, contrary to our findings, they did not find any statistical differences for *LHX2* and *LHX3* genes. Furthermore, the upregulation of *ZFPM1* was revealed in molecular high-risk patients with cytogenetically normal acute myeloid leukemia⁹⁴, yet its diagnostic value in CRC has not fully been confirmed⁹⁵. *SEMA7A* is also one of our selected hub genes that play a key role in several cancers including pancreatic, breast, and lung cancers^{69,96–98}. However, there has been less attention on the role of *SEMA7A* in CRC. Further investigation is required on our flagged DMGs.

Although there are many mechanisms that drive CRC, only a handful of them have been discovered in past studies. As researchers continue to genotype large panels of CRC tumors, it can be expected that additional new pathways of CRC carcinogenesis will be revealed. *SOST*, an identified hub gene in our study, plays a vital role in inhibiting the Wnt signaling pathway by binding to the Wnt co-receptor, LRP5/6, and preventing its activation⁹⁹. Therefore, decreased *SOST* expression could lead to an increase in Wnt signaling, promoting CRC cell proliferation, migration, and survival. Another identified hub gene is *TLE4* which is involved in the negative regulation of the canonical Wnt signaling pathway. Only a few investigations provided evidence of *TLE4* upregulation in CRC biopsies, partially through regulation of the JNK/c-Jun pathway⁷³. Moreover, recent studies that focus on the NFAT signaling pathway showed a promising strategy for CRC treatment⁶⁴.

Heterogeneity is one of the key features of genomic data. Specifically, there is evidence of the heterogeneity of DMG effects on the survival of CRC patients in the literature and in our dataset. The finite mixture of the AFT regression model is a plausible method to uncover such intangible heterogeneity. Our analysis suggested a mixture of two-component mixture of the AFT regression model in which patients were separated into two subgroups based on their vital status. In this model, almost all of the deceased patients were classified into the

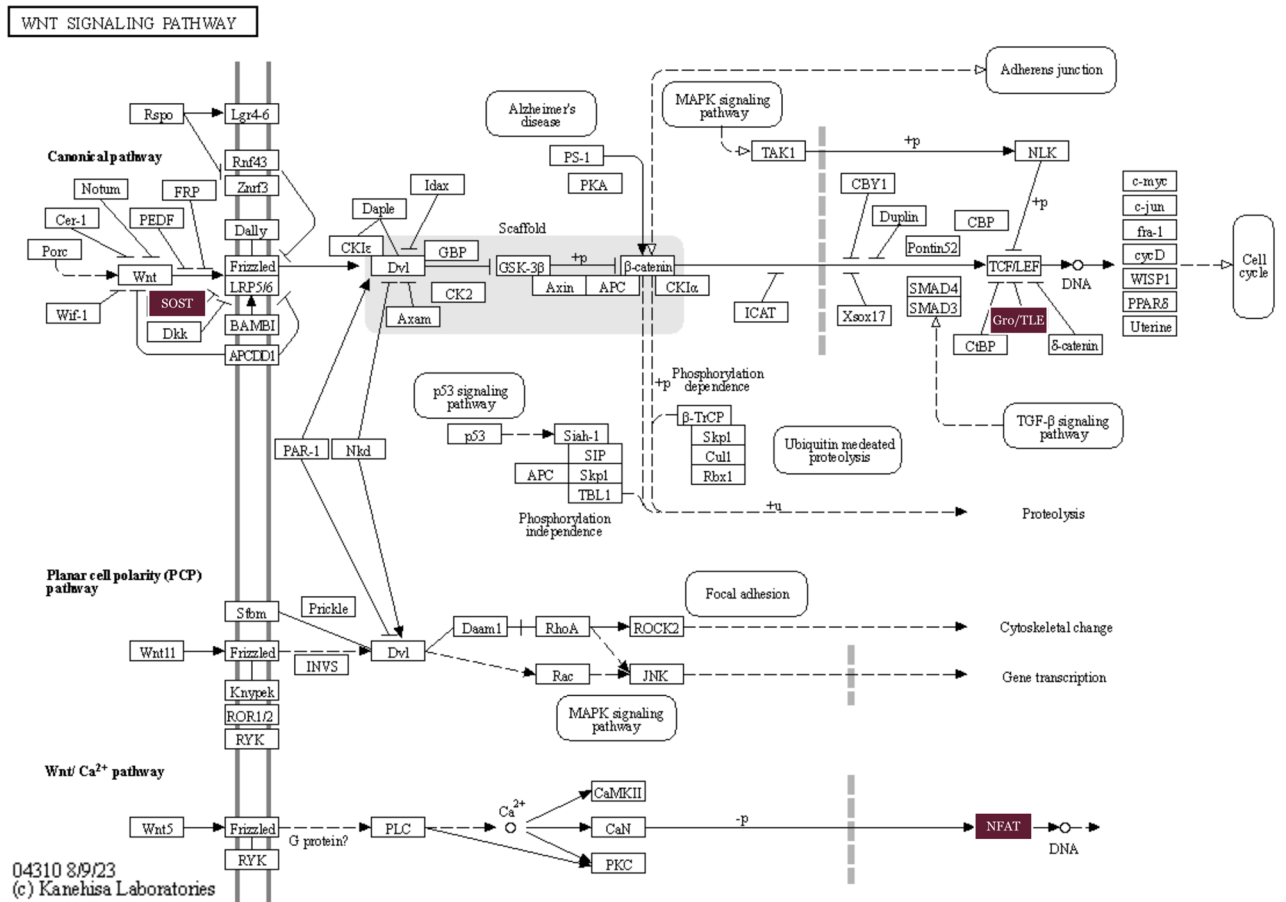


Figure 8. Wnt signaling pathway. The identified genes *SOST*, *Gro/TLE*, and *NFAT* are highlighted (R Shiny package ShinyGO, v0.77 & <https://www.kegg.jp/pathway/hsa04310>).

most aggressive form of the disease (Component 1). In Component 1, 83 DMGs including *NMNAT2*, *ZFP42*, *NPAS2*, *MYLK3*, *NUDT13*, *KIRREL3*, and *FKBP6* had an effect on the survival time of the patients. The relation between some of these DMGs and survival time has been previously reported¹⁰⁰. On the other hand, there are a few discoveries regarding other genes. For instance, significantly higher expression of *NMNAT2* in CRC tissues compared to normal ones have been found, yet this gene was not a prognostic factor for overall survival¹⁰¹. Note that, while the hub genes *SOST*, *NEATC1*, and *TLE4* were associated with survival in the univariate Cox model, they were only associated with survival time in the most aggressive form of the disease in our study.

Our study does not exclusively depend on bioinformatics analysis, as we have employed several statistical and machine learning analyses. These include modeling methylation profiles, identifying DMCs via *DMCHMM*, conducting statistical tests, performing multiple validation analyses, and applying statistical learning algorithms to survival times via *fmr.s*. One of the advantages of the *DMCHMM* method is that it does not require a large number of samples or matched samples, as it is highly flexible and can accommodate various experimental designs. It demonstrates significant power, particularly when dealing with moderate to low sample sizes.

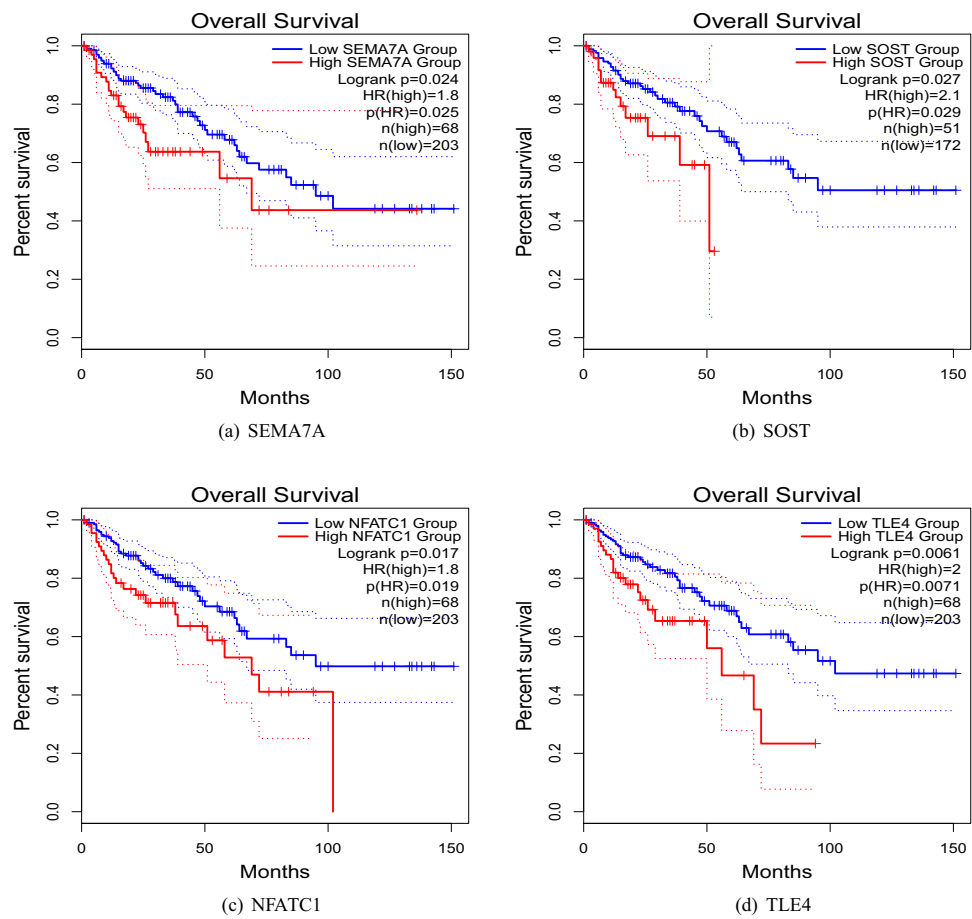


Figure 9. Overall survival of CRC patients stratified by their hub gene expression levels (<http://gepia2021.cancer-pku.cn>).

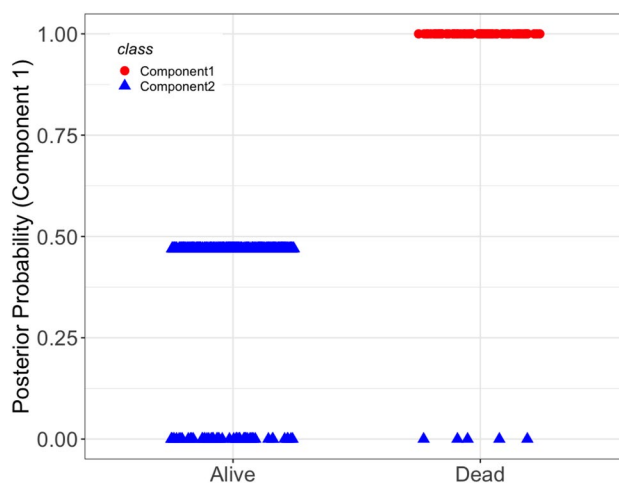


Figure 10. Posterior probability of CRC patients belonging to Component 1 separated for alive and deceased groups (R package *fmr*s, v2.0.1).

Gene	β_1	β_2	Gene	β_1	β_2	Gene	β_1	β_2
<i>NMI</i>	-27.2	0.0	<i>SIX6</i>	0.0	0.0	<i>FOXF2</i>	-12.1	0.0
<i>NCOA4</i>	-13.7	96,804.5	<i>FOXP2</i>	12.6	-101,775.9	<i>GIPR</i>	-19.0	0.0
<i>ANKMY1</i>	-32.6	0.0	<i>TNFSF9</i>	-14.7	0.0	<i>UCKL1</i>	-45.0	0.0
<i>ST6GAL2</i>	6.2	0.0	<i>CLDN3</i>	-2.1	21,941.6	<i>AMOTL1</i>	0.0	0.0
<i>PSMG3</i>	-12.4	-28,758.9	<i>DDX46</i>	40.0	0.0	<i>GMPS</i>	-6.8	0.0
<i>FAR2</i>	-21.6	0.0	<i>ZFAT</i>	0.0	0.0	<i>ADCY10</i>	0.0	0.0
<i>MPPED2</i>	-14.7	0.0	<i>OR5M1</i>	-6.0	0.0	<i>GPM6A</i>	-18.6	0.0
<i>GTF2IRD1</i>	-14.5	0.0	<i>PHACTR3</i>	6.3	0.0	<i>PFKP</i>	2.6	0.0
<i>FKBP6</i>	-11.9	0.0	<i>KRTAP13-4</i>	4.7	-15,847.1	<i>C14orf39</i>	2.4	-15,364.8
<i>SNORD109B</i>	-6.5	0.0	<i>LOC400940</i>	-6.6	70,576.5	<i>KCNK10</i>	0.0	0.0
<i>HLA-F</i>	0.0	0.0	<i>LRTM1</i>	-13.4	-50,609.5	<i>STK32B</i>	18.4	0.0
<i>AKAP9</i>	7.1	0.0	<i>NPAS2</i>	125.0	0.0	<i>IL1A</i>	13.3	0.0
<i>SEMA4F</i>	-21.3	0.0	<i>AXIN2</i>	24.3	0.0	<i>KRTAP20-1</i>	5.0	0.0
<i>RPL23P8</i>	18.1	0.0	<i>NKX2-3</i>	0.0	-13,689.5	<i>KIRREL2</i>	-13.1	0.0
<i>CHI3L1</i>	4.6	0.0	<i>NT5M</i>	18.8	0.0	<i>C1D</i>	-28.8	0.0
<i>NCAN</i>	3.7	151,828.5	<i>MECOM</i>	44.5	0.0	<i>EGR2</i>	54.7	0.0
<i>CLEC5A</i>	-10.4	0.0	<i>LUZP6</i>	-73.9	0.0	<i>PDF</i>	-1.4	8045.7
<i>TRPS1</i>	-16.7	0.0	<i>FLJ16779</i>	0.0	-87,806.8	<i>KCNQ3</i>	23.4	0.0
<i>CMKLR1</i>	18.1	0.0	<i>SLC25A24</i>	-6.0	-77,923.0	<i>CCR5</i>	-20.3	0.0
<i>GABRA4</i>	-6.2	0.0	<i>C1QTNF7</i>	-10.6	0.0	<i>COL4A3</i>	0.0	34,903.1
<i>OR5AS1</i>	-39.6	0.0	<i>MTNR1B</i>	11.7	0.0	<i>TFAP2C</i>	7.9	0.0
<i>MMP2</i>	0.0	0.0	<i>NMNAT2</i>	-12.0	0.0	<i>GNG2</i>	7.1	0.0
<i>AKAP12</i>	6.6	0.0	<i>BCKDK</i>	0.0	0.0	<i>OC90</i>	0.8	80,377.8
<i>PSD2</i>	5.4	-82,538.6	<i>ZFP42</i>	-13.8	0.0	<i>LHFPL2</i>	21.5	0.0
<i>FGFR1</i>	14.0	0.0	<i>CALB1</i>	-5.9	0.0	<i>STAU2</i>	0.0	0.0
<i>KIRREL3</i>	-10.3	0.0	<i>TCHH</i>	-17.8	0.0	<i>OLFM3</i>	10.3	0.0
<i>HECA</i>	-6.8	0.0	<i>MAPT</i>	-14.1	0.0	<i>SLTM</i>	-133.5	0.0
<i>MT1A</i>	0.0	125,763.0	<i>SYDE1</i>	4.2	-364,254.7	<i>NOC4L</i>	0.0	0.0
<i>NUDT13</i>	-7.3	0.0	<i>RNASE3</i>	7.0	0.0	<i>CNDP2</i>	0.0	0.0
<i>STON1-GTF2A1L</i>	-21.3	0.0	<i>PLCD3</i>	58.7	0.0	<i>NFATC1</i>	-20.3	0.0
<i>LBP</i>	-7.7	0.0	<i>MAP1LC3A</i>	5.8	0.0	<i>SEMA7A</i>	21.4	0.0
<i>MYLK3</i>	21.9	0.0	<i>CROCC</i>	18.2	0.0	<i>SOST</i>	-2.5	0.0
<i>RFPL4B</i>	0.0	0.0	<i>OPCML</i>	21.4	0.0	<i>TLE4</i>	-7.2	0.0

Table 3. Estimated DMG effects in the two-component mixture of accelerated failure time regression model in the CRC data.

Gene	Information	Summary
CLDN3	Description	Claudin 3
	Predicted location	Membrane
	Protein class	Cancer-related genes, Disease-related genes, Potential drug targets, Transporters
	Cell line specificity	Cancer enhanced (CRC)
	Pathway	Cell adhesion tight junctions, Cell adhesion endothelial cell contacts by junctional mechanisms
	Function	Contributes to the closure of intercellular gaps within tight junctions through calcium-independent cell adhesion.
	Cancer	Tends to be down-regulated in primary CRC samples and can predict prognosis in CMS2 or CMS3 CRC subtypes.
	Reference	Perez et al. ⁶¹ , Cherradi et al. ⁶²
NFATC1	Description	Nuclear Factor of Activated T Cells 1
	Predicted location	Intracellular
	Protein class	Transcription factors
	Cell line specificity	Cancer enhanced (Lymphoma)
	Pathway	Activation of cAMP-dependent PKA, Activation of PKA through GPCR, APRIL pathway, BAFF in B-Cell signaling, cAMP pathway
	Function	Contributes to the inducible expression of cytokine genes in T-cells, influencing the transcription of genes like IL-2 and IL-4. It also affects gene expression in embryonic cardiac cells, and plays a role in T-lymphocyte activation, proliferation, differentiation, and programmed cell death.
	Cancer	Activates the transcription of SNAIL1, facilitating EMT and CRC metastasis. It's an immune-related prognostic risk factor for CRC immunotherapy.
	Reference	Chuvpilo et al. ⁶³ , Shen et al. ⁶⁴ , Wu et al. ⁶⁵
AXIN2	Description	Axin 2
	Predicted location	Intracellular
	Protein class	Cancer-related genes, Disease-related genes, Plasma proteins
	Cell line specificity	Cancer enhanced (CRC, Gastric cancer)
	Pathway	Wnt signaling pathway, Cytoskeleton remodeling reverse signaling by ephrin B
	Function	Plays a role in stabilizing beta-catenin within the Wnt signaling pathway, similar to mouse conductin and rat axil in rodents.
	Cancer	AXIN1/2 alterations may be key defects in some cancers including CRC and hepatocellular carcinoma.
	Reference	Mazzoni et al. ⁶⁶
SEMA7A	Description	Semaphorin 7A
	Predicted location	Membrane
	Protein class	Disease-related genes
	Cell line specificity	Low cancer specificity
	Pathway	Axon guidance, Developmental biology, Nervous system development, Other semaphorin interactions, Semaphorin interactions
	Function	Has a significant role in integrin-mediated signaling, governing cell migration and immune reactions. Facilitates the assembly of focal adhesion complexes, triggers the activation of protein kinase PTK2/FAK1, leading to MAPK1 and MAPK3 phosphorylation.
	Cancer	Associated with Breast, Lung, and Pancreatic cancers.
	Reference	Mastrantonio et al. ⁶⁷ , Fijneman et al. ⁶⁸ , Liu et al. ⁶⁹
UCKL1	Description	Uridine-Cytidine Kinase 1 Like 1
	Predicted location	Intracellular
	Protein class	Enzymes, Metabolic proteins
	Cell line specificity	Low cancer specificity
	Pathway	Pyrimidine metabolism
	Function	Encodes a uridine kinase, converting uridine into uridine monophosphate. Its ubiquitination increases with natural killer lytic-associated molecule presence, resulting in protein degradation. A potential therapeutic target for inhibiting tumor growth and metastasis.
	Cancer	A candidate gene in CRC.
	Reference	Long et al. ⁷⁰ , Matchett et al. ⁷¹
ANKMY1	Description	Ankyrin Repeat and MYND Domain containing 1
	Predicted location	Intracellular
	Protein class	-
	Cell line specificity	Low cancer specificity
	Pathways	WP5224 pathway
	Function	Predicted to enable metal ion binding activity.
	Cancer	Associated with Osteosarcoma.
	Reference	Wang et al. ⁷²
Continued		

Gene	Information	Summary
TLE4	Description	TLE family member 4, transcriptional corepressor
	Predicted location	Intracellular
	Protein class	–
	Cell line specificity	Low cancer specificity
	Pathway	Wnt signaling pathway, Development Notch signaling pathway
	Function	Transcriptional corepressor that binds to various transcription factors. Inhibits transcriptional activation by PAX5, CTNNB1, and TCF family members in the Wnt signaling pathway.
	Cancer	Overexpression may play a role in CRC development and progression, partly through the JNK/c-Jun pathway. It is a candidate for risk stratification of cancer recurrence after curative resection of early-stage CRC.
	Reference	Wang et al. ⁷³ , Yu et al. ⁷⁴
EGR2	Description	Early Growth Response 2
	Predicted location	Localized to the Nucleoplasm
	Protein class	Disease related genes, Transcription factors
	Cell line specificity	Cancer enhanced (Lymphoma)
	Pathway	Activation of anterior HOX genes in hindbrain development during early embryogenesis, Activation of HOX genes during differentiation
	Function	Mutations linked to Charcot-Marie-Tooth disease type 1D, Charcot-Marie-Tooth disease type 4E, and Dejerine-Sottas syndrome.
	Cancer	Targeting EGR2 may provide a therapeutic strategy to eliminate colon cancer stem cells and block nervous system-driven disease progression through differentiation.
	Reference	Regan et al. ⁷⁵
SLTM	Description	SAFB Like Transcription Modulator
	Predicted location	Intracellular
	Protein class	–
	Cell line specificity	Low cancer specificity
	Pathway	–
	Function	Hypothesized to play a role in mRNA processing regulation and RNA polymerase II-mediated transcription regulation.
	Cancer	Up-regulated in dextran sulfate sodium treated colon mucosa.
	Reference	De Robertis et al. ⁷⁶
PLCD3	Description	Phospholipase C Delta 3
	Predicted location	Intracellular
	Protein class	Enzymes, Metabolic proteins, Plasma proteins
	Cell line specificity	Low cancer specificity
	Pathway	Wnt signaling pathway
	Function	Crucial for trophoblast and placental development, possibly contributing to cytokinesis by cleavage furrow PIP2 hydrolysis. Controls neurite outgrowth by suppressing RhoA/Rho kinase signaling.
	Cancer	Down-regulation of Phosphatidylinositol signaling system pathway in CRC mucosa.
	Reference	Danielsen et al. ⁷⁷
MAPT	Description	Microtubule Associated Protein Tau
	Predicted location	Intracellular
	Protein class	Disease related genes, FDA approved drug targets, Plasma proteins
	Cell line specificity	Group enriched (Bone cancer, Neuroblastoma)
	Pathway	AMPK signaling pathway, P38 MAPK signaling pathway
	Function	Promotes microtubule assembly and stability, maintaining neuronal polarity. Binds axonal microtubules and neural plasma membrane components, acting as a link between them. Its localization within the cell body helps define axonal polarity.
	Cancer	Hyper-methylation in MAPT is associated with poor prognosis in stage II CRC patients.
	Reference	Sandberg et al. ⁷⁸ , Wang et al. ⁷⁹
FOXF2	Description	Forkhead box F2
	Predicted location	Intracellular
	Protein class	Transcription factors
	Cell line specificity	Low cancer specificity
	Pathway	–
	Function	Among human counterparts of Drosophila melanogaster forkhead transcription factor. Expresses in the lung and placenta, activating transcription of several lung-specific genes.
	Cancer	Regulates PRUNE2 transcription in CRC pathogenesis and is hyper-methylated in CRC samples.
	Reference	Li et al. ⁸⁰ , Hauptman et al. ⁸¹
Continued		

Gene	Information	Summary
TNFSF9	Description	TNF superfamily member 9
	Predicted location	Intracellular
	Protein class	Plasma proteins
	Cell line specificity	Cancer enhanced (Kidney cancer)
	Pathway	Cytokine signaling in immune system, TNFR2 non-canonical NF-kB pathway
	Function	Cytokine in the TNF ligand family, acting as a bidirectional signal transducer with TNFRSF9/4-1BB. Key role in antigen presentation, cytotoxic T cell generation, and T lymphocyte activation and proliferation.
	Cancer	Critical role in liver homing for metastatic colon cancer.
	Reference	Barderas et al. ⁸²

Table 4. Bioinformatics and biological information of selected DMGs related to survival in colorectal cancer patients.

Data availability

In this study methylation profiling datasets with accession numbers GSE53051, GSE77718, GSE101764, GSE42752, and GSE48684 were obtained from Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>), of the National Center for Biotechnology Information (NCBI). Additional DNA methylation datasets and expression profiles of CRC patients (TCGA-COAD, TCGA-READ, TCGA-SARC projects) were obtained from The Cancer Genome Atlas (TCGA, <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>), of the National Cancer Institute (NCI). Our SureSelectXT Human Methyl-Seq dataset on methylation profiles of 6 patients with adenocarcinoma of CRC and 6 normal males is obtained from 'Reza Radiotherapy and Oncology Center' in Iran and is available upon request.

Received: 19 May 2023; Accepted: 13 November 2023

Published online: 13 December 2023

References

- Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
- Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol. Mech. Dis.* **6**, 479–507 (2011).
- Andrew, A. *et al.* Risk factors for diagnosis of colorectal cancer at a late stage: A population-based study. *J. Gen. Intern. Med.* **33**, 2100–2105 (2018).
- Das, P. & Singal, R. DNA methylation and cancer. *J. Clin. Oncol.* **22**, 4632–4642 (2004).
- Moore, L., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).
- Ashktorab, H. & Brim, H. DNA methylation and colorectal cancer. *Curr. Colorectal Cancer Rep.* **10**, 425–430 (2014).
- Grady, W. Epigenetic events in the colorectum and in colon cancer. *Biochem. Soc. Trans.* **33**, 684–688 (2005).
- Lam, K. *et al.* DNA methylation-based biomarkers in colorectal cancer: A systematic review. *Biochim. Biophys. Acta* **1866**, 106–120 (2016).
- Payne, S. R. From discovery to the clinic: The novel DNA methylation biomarker m SEPT9 for the detection of colorectal cancer in blood. *Epigenomics* **2**, 575–585 (2010).
- Imperiale, T. *et al.* Multitarget stool DNA testing for colorectal-cancer screening. *N. Engl. J. Med.* **370**, 1287–1297 (2014).
- Mathers, J., Strathdee, G. & Relton, C. Induction of epigenetic alterations by dietary and other environmental factors. *Adv. Genet.* **71**, 3–39 (2010).
- Issa, J.-P. *et al.* Methylation of the oestrogen receptor CpG island links aging and neoplasia in human colon. *Nat. Genet.* **7**, 536–540 (1994).
- Barrow, T. M. *et al.* Smoking is associated with hypermethylation of the APC 1A promoter in colorectal cancer: The ColoCare Study. *J. Pathol.* **243**, 366–375 (2017).
- Shokoohi, F. *et al.* A hidden Markov model for identifying differentially methylated sites in bisulfite sequencing data. *Biometrics* **75**, 210–221 (2019).
- Al-Sohaily, S. *et al.* Molecular pathways in colorectal cancer. *J. Gastroenterol. Hepatol.* **27**, 1423–1431 (2012).
- Ilyas, M., Straub, J., Tomlinson, I. & Bodmer, W. Genetic pathways in colorectal and other cancers. *Eur. J. Cancer* **35**, 1986–2002 (1999).
- Behrens, J. The role of the Wnt signaling pathway in colorectal tumorigenesis. *Biochem. Soc. Trans.* **33**, 672–675 (2005).
- Fang, J. & Richardson, B. The MAPK signaling pathways and colorectal cancer. *Lancet Oncol.* **6**, 322–327 (2005).
- Markowitz, S. *et al.* Inactivation of the type II TGF- β receptor in colon cancer cells with microsatellite instability. *Science* **268**, 1336–1338 (1995).
- Levine, A. & Oren, M. The first 30 years of p53: Growing ever more complex. *Nat. Rev. Cancer* **9**, 749–758 (2009).
- Gong, B. *et al.* Identification of hub genes related to carcinogenesis and prognosis in colorectal cancer based on integrated bioinformatics. *Mediat. Inflamm.* **2020**, 1–11 (2020).
- Huang, H. *et al.* Integrative analysis of identifying methylation-driven genes signature predicts prognosis in colorectal carcinoma. *Front. Oncol.* **11**, 629860 (2021).
- Hu, J. *et al.* An eight-CpG-based methylation classifier for preoperative discriminating early and advanced-late stage of colorectal cancer. *Front. Genet.* **11**, 614160 (2021).
- Feng, Z., Liu, Z., Peng, K. & Wu, W. A prognostic model based on nine DNA methylation-driven genes predicts overall survival for colorectal cancer. *Front. Genet.* **12**, 2446 (2022).
- Long, J. *et al.* DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. *Theranostics* **9**, 7251 (2019).
- Shokoohi, F., Khalili, A., Asgharian, M. & Lin, S. Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates. *Ann. Appl. Stat.* **13**, 444 (2019).
- Welchowski, T., Zuber, V. & Schmid, M. Correlation-adjusted regression survival scores for high-dimensional variable selection. *Stat. Med.* **38**, 2413–2427 (2019).

28. Kerachian, M. A. *et al.* Crosstalk between dna methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor. *Sci. Rep.* **10**, 1–13 (2020).
29. Wingett, S. W. & Andrews, S. Fastq screen: A tool for multi-genome mapping and quality control. *F1000 Res.* **7**, 1–10 (2018).
30. Krueger, F. *Trim Galore!: A Wrapper Around Cutadapt and Fastqc to Consistently Apply Adapter and Quality Trimming to Fastq Files, with Extra Functionality for RRBS Data* (Babraham Institute, 2015).
31. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
32. Shokoohi, F. DMCHMM: Differentially methylated CpG using hidden Markov model. *Bioconductor* <https://doi.org/10.18129/B9.bioc.DMCHMM> (2023).
33. Timp, W. *et al.* Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.* **6**, 1–11 (2014).
34. McInnes, T. *et al.* Genome-wide methylation analysis identifies a core set of hypermethylated genes in CIMP-H colorectal cancer. *BMC Cancer* **17**, 1–11 (2017).
35. Naumov, V. A. *et al.* Genome-scale analysis of DNA methylation in colorectal cancer using Infinium human methylation 450 bead chips. *Epigenetics* **8**, 921–934 (2013).
36. Luo, Y. *et al.* Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* **147**, 418–429 (2014).
37. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
38. Tim Triche, J. FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes. R package version 2.2.0. (2014).
39. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
40. Sherman, B. T. *et al.* David: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).
41. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. Clusterprofiler: An r package for comparing biological themes among gene clusters. *Omic* **16**, 284–287 (2012).
43. Ulgen, E., Ozisik, O. & Sezerman, O. U. pathfinder: An r package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Front. Genet.* **10**, 858 (2019).
44. Ge, S. X., Jung, D. & Yao, R. Shinygo: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
45. Luo, W. & Brouwer, C. Pathview: An r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).
46. Network, C. G. A. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330 (2012).
47. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
48. Bordes, L. & Chauveau, D. Stochastic em algorithms for parametric and semiparametric mixture models for right-censored lifetime data. *Comput. Stat.* **31**, 1513–1538 (2016).
49. Matthaïos, D. *et al.* Methylation status of the apc and rass1a promoter in cell-free circulating dna and its prognostic role in patients with colorectal cancer. *Oncol. Lett.* **12**, 748–756 (2016).
50. Chen, S.-P. *et al.* β -catenin and k-ras mutations and rass1a promoter methylation in Taiwanese colorectal cancer patients. *Genet. Test. Mol. Biomark.* **16**, 1277–1281 (2012).
51. Muthamilselvan, S., Raghavendran, A. & Palaniappan, A. Stage-differentiated ensemble modeling of dna methylation landscapes uncovers salient biomarkers and prognostic signatures in colorectal cancer progression. *PLoS ONE* **17**, e0249151 (2022).
52. Joo, J. E. *et al.* Dna methylation signatures and the contribution of age-associated methylomic drift to carcinogenesis in early-onset colorectal cancer. *Cancers* **13**, 2589 (2021).
53. Shokoohi, F. Sparse estimation in finite mixture of accelerated failure time and mixture of regression models with R package fmrs. *N. Engl. J. Stat. Data Sci.* **1**, 1–18. <https://doi.org/10.51387/23-NEJSDS49> (2023).
54. Kolde, R. pheatmap: Pretty Heatmaps (2019). R package version 1.0.12.
55. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in r. *Bioinformatics* **30**, 2811–2812 (2014).
56. Dusa, A. venn: Draw Venn Diagrams (2022). R package version 1.1.1.
57. Gel, B. & Serra, E. karyoploter: An r / bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090. <https://doi.org/10.1093/bioinformatics/btx346> (2017).
58. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
59. Tang, Z. *et al.* Gepia: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).
60. Shokoohi, F. *fmrs: Variable Selection in Finite Mixture of AFT Regression and FMR*. *Bioconductor Open Source Software for Bioinformatics*. <https://doi.org/10.18129/B9.bioc.fmrs> (2023). R package version 2.0.1.
61. Pérez, A. G. *et al.* N-glycosylation and receptor tyrosine kinase signaling affect claudin-3 levels in colorectal cancer cells. *Oncol. Rep.* **44**, 1649–1661 (2020).
62. Cherradi, S., Martineau, P., Gongora, C. & Del Rio, M. Claudin gene expression profiles and clinical value in colorectal tumors classified according to their molecular subtype. *Cancer Manag. Res.* **1**, 1337–1348 (2019).
63. Chuvpilo, S. *et al.* Multiple nf-atc isoforms with individual transcriptional properties are synthesized in t lymphocytes. *J. Immunol.* **162**, 7294–7301 (1999).
64. Shen, T. *et al.* NFATc1 promotes epithelial-mesenchymal transition and facilitates colorectal cancer metastasis by targeting SNA11. *Exp. Cell Res.* **408**, 112854 (2021).
65. Wu, B. *et al.* Development of an immune infiltration-related eight-gene prognostic signature in colorectal cancer microenvironment. *BioMed. Res. Int.* **2020** (2020).
66. Mazzoni, S. M. & Fearon, E. R. Axin1 and axin2 variants in gastrointestinal cancers. *Cancer Lett.* **355**, 1–8 (2014).
67. Mastrantonio, R., You, H. & Tamagnone, L. Semaphorins as emerging clinical biomarkers and therapeutic targets in cancer. *Theranostics* **11**, 3262 (2021).
68. Fijneman, R. J. *et al.* Proximal fluid proteome profiling of mouse colon tumors reveals biomarkers for early diagnosis of human colorectal cancer. *Clin. Cancer Res.* **18**, 2613–2624 (2012).
69. Liu, Y., Guo, C., Li, F. & Wu, L. LncRNA LOXL1-AS1/miR-28-5p/SEMA7A axis facilitates pancreatic cancer progression. *Cell Biochem. Funct.* **38**, 58–65 (2020).
70. Long, N. P. *et al.* Novel biomarker candidates for colorectal cancer metastasis: A meta-analysis of in vitro studies. *Cancer Inform.* **15**, 40301 (2016).

71. Matchett, E. C., Ambrose, E. C. & Kornbluth, J. Characterization of uridine-cytidine kinase like-1 nucleoside kinase activity and its role in tumor growth. *Biochem. J.* **479**, 1149–1164 (2022).
72. Wang, F., Qin, G., Liu, J., Wang, X. & Ye, B. Integrated genome-wide methylation and expression analyses reveal key regulators in osteosarcoma. *Comput. Math. Methods Med.* **2020** (2020).
73. Wang, S.-Y. *et al.* TLE4 promotes colorectal cancer progression through activation of JNK/c-Jun signaling pathway. *Oncotarget* **7**, 2878 (2016).
74. Yu, H. *et al.* DNA methylation profile in CpG-depleted regions uncovers a high-risk subtype of early-stage colorectal cancer. *JNCI J. Natl. Cancer Inst.* **115**, 52–61 (2023).
75. Regan, J. L. *et al.* Identification of a neural development gene expression signature in colon cancer stem cells reveals a role for egr2 in tumorigenesis. *Science* **25**, 104498 (2022).
76. De Robertis, M. *et al.* Novel insights into notum and glypicans regulation in colorectal cancer. *Oncotarget* **6**, 41237 (2015).
77. Danielsen, S. A. *et al.* Phospholipase c isozymes are deregulated in colorectal cancer—insights gained from gene set enrichment analysis of the transcriptome. *PLoS ONE* **6**, e24419 (2011).
78. Sandberg, A. *et al.* Fibrillation and molecular characteristics are coherent with clinical and pathological features of 4-repeat tauopathy caused by mapt variant g273r. *Neurobiol. Dis.* **146**, 105079 (2020).
79. Wang, C. *et al.* Mapt promoter cpG island hypermethylation is associated with poor prognosis in patients with stage ii colorectal cancer. *Cancer Manag. Res.* **1**, 7337–7343 (2019).
80. Li, T., Huang, S., Yan, W., Zhang, Y. & Guo, Q. Foxf2 regulates prune2 transcription in the pathogenesis of colorectal cancer. *Technol. Cancer Res. Treatm.* **21**, 15330338221118716 (2022).
81. Hauptman, N., Jevšinek Skok, D., Spasovska, E., Boštjančič, E. & Glavač, D. Genes cep55, foxd3, foxf2, gnao1, gria4, and kcna5 as potential diagnostic biomarkers in colorectal cancer. *BMC Med. Genomics* **12**, 1–17 (2019).
82. Barderas, R. *et al.* In-depth characterization of the secretome of colorectal cancer metastatic cells identifies key proteins in cell adhesion, migration, and invasion. *Mol. Cell. Proteom.* **12**, 1602–1620 (2013).
83. Imperiale, T. *et al.* Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. *N. Engl. J. Med.* **351**, 2704–2714 (2004).
84. Mueller, D. & Györfy, B. DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer. *Biochim. Biophys. Acta* **1877**, 1–12 (2022).
85. Gutierrez, A., Demond, H., Brebi, P. & Ili, C. Novel methylation biomarkers for colorectal cancer prognosis. *Biomolecules* **11**, 1722 (2021).
86. He, Y. *et al.* NK homeobox 2.2 functions as tumor suppressor in colorectal cancer due to DNA methylation. *J. Cancer* **11**, 4791 (2020).
87. Küçükköse, E. *et al.* KIT promotes tumor stroma formation and counteracts tumor-suppressive TGF β signaling in colorectal cancer. *Cell Death Dis.* **13**, 617 (2022).
88. Xiao, Q. *et al.* MEF2A transcriptionally upregulates the expression of ZEB2 and CTNBN1 in colorectal cancer to promote tumor progression. *Oncogene* **40**, 3364–3377 (2021).
89. Pellatt, A. J. *et al.* The TGF β -signaling pathway and colorectal cancer: Associations between dysregulated genes and miRNAs. *J. Transl. Med.* **16**, 1–22 (2018).
90. Li, F. *et al.* Transcriptomic landscape of sodium butyrate-induced growth inhibition of human colorectal cancer organoids. *Mol. Omics* **18**, 754–764 (2022).
91. Rotte, A. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. *J. Exp. Clin. Cancer Res.* **38**, 1–12 (2019).
92. Yuan, C., Wu, C., Xue, R., Jin, C. & Zheng, C. Suppression of human colon tumor by EERAC through regulating Notch/DLL4/Hes pathway inhibiting angiogenesis in vivo. *J. Cancer* **12**, 5914 (2021).
93. Cha, N. *et al.* Oncogenicity of LHX4 in colorectal cancer through Wnt/ β -catenin/TCF4 cascade. *Tumor Biol.* **35**, 10319–10324 (2014).
94. Marcucci, G. *et al.* Prognostic significance of, and gene and microRNA expression signatures associated with, CEBPA mutations in cytogenetically normal acute myeloid leukemia with high-risk molecular features: A Cancer and Leukemia Group B Study. *J. Clin. Oncol.* **26**, 5078 (2008).
95. Ge, W. *et al.* A novel 4-gene prognostic signature for hypermutated colorectal cancer. *Cancer Manag. Res.* **11**, 1985 (2019).
96. Rehman, M. & Tamagnone, L. Semaphorins in cancer: Biological mechanisms and therapeutic approaches. *Semin. Cell Dev. Biol.* **24**, 179–189 (2013).
97. Crump, L. S. *et al.* Hormonal regulation of Semaphorin 7a in ER+ breast cancer drives therapeutic resistance. *Cancer Res.* **81**, 187–198 (2021).
98. Kinehara, Y. *et al.* Semaphorin 7A promotes EGFR-TKI resistance in EGFR mutant lung adenocarcinoma cells. *JCI Insight* **3**, 1–17 (2018).
99. Katoh, M. & Katoh, M. Molecular genetics and targeted therapy of Wnt-related human diseases. *Int. J. Mol. Med.* **40**, 587–606 (2017).
100. Yang, S.-F., Xu, M., Yang, H.-Y., Li, P.-Q. & Chi, X.-F. Expression of circadian gene NPAS2 in colorectal cancer and its prognostic significance. *Nan Fang Yi Ke Da Xue Xue Bao* **36**, 714–718 (2016).
101. Cui, C. *et al.* Nicotinamide mononucleotide adenylyl transferase 2: A promising diagnostic and therapeutic target for colorectal cancer. *BioMed Res. Int.* **2016**, 1–8 (2016).

Acknowledgements

The authors would like to express their gratitude to the editor, the associate editor, and the two anonymous reviewers for their thoughtful and constructive comments, which significantly improved the manuscript. The authors also wish to extend their thanks to Prof. Kazem Taghva, the Chair of the Department of Computer Science, and Prof. Martin Schiller, the Director of the Nevada Institute for Personalized Medicine (NIPM) at the University of Nevada, Las Vegas (UNLV), for their financial support and assistance. Furthermore, the authors appreciate the generosity of the Reza Radiotherapy and Oncology Center in Iran, Mashhad University of Medical Sciences, for sharing the CRC dataset. The publication fees for this article were supported by the UNLV University Libraries Open Article Fund.

Author contributions

S.H.K. and F.S. analyzed the data and wrote the paper. H.E. and M.A.K. edited the paper. All authors read and approved the final manuscript.

Funding

Farhad Shokoohi is supported by Start-up Grant number PG18929 and 'In Support of Research Scholar' Grant Number PG18494, University of Nevada, Las Vegas. This research is partially supported by NIPM, the Department of Computer Science at UNLV, and the Center of Biomedical Research Excellence through COBRE Pilot Grant Number P20GM121325.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-47377-1>.

Correspondence and requests for materials should be addressed to F.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023