**Genome-wide DNA methylation profiles of low- and high-grade adenoma reveals potential early diagnosis biomarkers for colorectal carcinoma**

Jian Fan1,4 #, Jun Li2 #, Shicheng Guo3 # ,Chengcheng Tao1, Haikun Zhang1,4, Wenjing Wang2, 5**,** Ying Zhang1, Steven J Schrodi3,6,7, Dake Zhang1\*, Shigang Ding2\*, Changqing Zeng1\*.

1，Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

2，Department of Gastroenterology, Peking University Third Hospital, Beijing 100191, China

3，Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA

4，University of Chinese Academy of Sciences, Beijing 100049, China

5，Center for Oral and Systemic Health, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA

6. Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, 53706, USA

7. Department of Medical Genetics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA

# These authors contributed equally to this work; \* Corresponding Author

**Shigang Ding, M.D.**

Department of Gastroenterology, Peking University Third Hospital, Beijing 100191, China

Tel: (010)8226-69905

Email: dingshigang222@163.com

**Dake Zhang, Ph.D.**

Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics

Chinese Academy of Sciences, Beijing, 100101,

Tel: (010) 8409-7566

Email: [zhangdk@big.ac.cn](mailto:zhangdk@big.ac.cn)

**Changqing Zeng, Ph.D.**

Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics

Chinese Academy of Sciences, Beijing, 100101,

Tel: (010) 8409-7818

Email: [czeng@big.ac.cn](mailto:czeng@big.ac.cn)

## **Abstract**

DNA methylation abnormalities is a hallmark of human cancers and have demonstrated inherent potential as as candidate biomarkers for early diagnosis to human cancers. In this study, we applied Illumina methylation 450K beadchip to identify colorectal cancer early diagnostic biomarkers based on pre-colorectal cancer samples including low-grade and high-grade adenoma (N=60). We identified that 209 and 8,692 CpG sites were significantly hyper-methylated in the low-grade adenoma (LGA) and high-grade adenoma (HGA), respectively. Pathway analysis identified that nervous system is significantly associated with early adenoma development. Together with GEO and TCGA dataset (N=833), integration analysis revealed that DNA methylation in the promoter of *ADHFE1* is a most potential early diagnostic biomarker for colorectal adenoma and cancer (SEN=0.96, SPE=0.95, AUC=0.97).

**Key words:**

DNA methylation, Adenoma, Colorectal cancer, biomarker

## **Background**

Colorectal cancer (CRC) is the third leading cause of cancer-related deaths 1, 2. Evidence shows not only genetic mutations but also epigenetic alterations progressively accumulatein the tumor genome during the occurrence of human cancers. While DNA methylation plays important roles in embryonic development and tissue differentiation, abnormal hyper-methylation or hypo-methylation has also been associated with abnormal tissue differentiation. Altered methylation has been observed in the promoter regions of tumor suppressor genes and miRNA in almost all cancer types 3, 4. In recent decades, detection of altered DNA methylation has been widely applied as a potential candidate cancer biomarker 5. For example altered methylation patterns have been detected with hepatic disease progression in the context of hepatitis, cirrhosis and hepatocellular carcinoma (HCC) 6, 7. Moreover, recent evidence demonstrated that cell free DNA (cfDNA) methylation can be used for early cancer diagnosis and tissue-of-origin mapping 3.

Abnormal alterations of DNA methylation have been recognized as an important event of cancer development and early research hypothesized methylation changes may be detected at all stages of cancer. Global hypo-methylation arises early, and was linked to chromosomal instability and loss of imprinting 8, 9. Generally, during cancer development, hundreds of genes are [silenced or activated](https://en.wikipedia.org/wiki/Regulation_of_transcription_in_cancer#Transcription_silencing/activation_in_cancers) 10, 11, 12. Although silencing of some genes in cancers occurs by mutation, a large proportion of carcinogenic gene silencing is a result of altered DNA methylation. DNA methylation causing silencing in cancer typically occurs at multiple CpG sites in the [CpG islands](https://en.wikipedia.org/wiki/CpG_site#CpG_island) that are present in the promoters of protein-coding genes 13.

Although extensive epigenetic alterations have been defined over the past years, CRC is still not well understood at the molecular level. Against a background of whole genome hypo-methylation, gene-specific promoter hyper-methylation was found to promote CRC by down-regulating the expression of key tumor suppressor genes, such as *CDKN2A*, *MLH1*, and *CDH1* 14, 15, 16. CRC is a heterogeneous disease, which typically originates from a [benign tumor](https://en.wikipedia.org/wiki/Adenoma), often in the form of an adenoma, eventually progressing to a malignant cancer within a temporal window that may exceed 10 years 17. Because CRC exceeds many other cancers in both incidence and mortality, capacity to detect and monitor molecular changes during the colorectal adenoma (CA) stage provides an excellent opportunity to prevent cancer progression and improve survival outcomes. While a large number of studies have focused on CRC, a subset have focused specifically on the adenoma as an intermediate stage which required more specific molecular definition. . Notably, colorectal adenoma has two different pathologic stages: low-grade adenoma (LGA) and high-grade adenoma (HGA). Our research previously compared and defined differences in whole-genome DNA methylation across these two stages that had not been previously reported 18. We hypothesized that these alterations in LGA methylation represented candidates as potential early diagnostic biomarkers. We further posit that comprehensive understanding to the genome-wide DNA methylation profile for the early stage pre-cancerous lesions (LGA and HGA) will provide important resources early diagnosis and candidate biomarkers for potential oncogenic progression.

In this study, we characterized adenoma at the level of in both stages and conducted genome-wide DNA methylation array on 18 LGA, 22 HGA and 20 normal tissues samples. Dynamic DNA methylation changes were identified for LGA and HGA and we found the methylation change appeared in LGA will strengthen or maintain in LGA and caner. We conducted enrichment analysis of differential methylation regions (DMRs) to investigate the potential influence of DNA methylation on functional difference in adenoma initiation and development stages. Moreover, we separated different methylation sites (DMSs) between LGA and normal into hyper-DMS and hypo-DMS, and evaluated their respective performance for the CA and CRC prediction. Meanwhile, we collected genome-wide DNA methylation profile of 833 samples from public database to validate our findings. Finally, we described one functional methylation biomarker, *ADHFE1*, for colorectal adenoma and cancer.

## **Results**

#### Landscape of DNA methylation of pre-cancerous benign lesion

We profiled DNA methylation at single-base level for 18 low-grade adenoma (LGA), 22 high-grade colorectal adenoma (HGA) and 20 normal tissue (**Method**). We found the significant genome-wide DNA methylation difference among normal, low and high-grade adenoma (**Figure 1A** and **1B**). Compared with the normal tissue, low-grade adenoma had genome wide hypo-methylation (P = 5.2x10-5, rank sum test), and global methylation level became even lower in high-grade adenoma (P = 3.7x10-6, compared with normal, rank sum test, **Figure 1C**). Methylation levels of all target sites in the array demonstrated the known bimodal distribution in normal, LGA and HGA (**Figure 1D**), and the amount of fully methylated sites decreased with increased degree of malignancy of lesions (right peak, **Figure1D**; **Figure 1E**). Almost all DMSs in LGA compared with normal tissues demonstrated at least equivalent or higher methylation levels, compared to those observed for HGA and cancer. The 209 significantly hyper-methylated sites in LGA showed further methylation in 22 HGA and 504 cancer samples collected from the public databases (**Figure 1F**) and hypo-DMSs showed diametric tendency (**Figure S2**), demonstrating that DNA demethylation may occur in very early in precancerous lesions.Notably , over 60% DMRs were hypo-methylated observed in both LGA (71.4%, 314/440) and HGA (61.9%, 4,213/6,805) in comparison with normal tissues (**Figure 1G, Table S1-2**). However, with LGA as the reference, most DMRs observed in HGA were hyper-methylated (76.0%, 660/868) (**Figure 1G, Table S3**). This suggested genome wide demethylation may dominate the malignant transformation course of tissue cells, but hyper-methylation sites may contribute more to the distinct malignancy of these lesions. In addition, there was limited overlap between genes with DMRs in LGA compared to normal tissues and those compared to HGA, suggesting different epigenetic processes (**Figure 1H**)19.

#### Nervous system was associated with adenoma development

Enrichment analysis of 603 DMRs between high-grade adenoma and low-grade adenoma located on, and most highly enriched functional terms included the nervous system and those associated with signal transduction (**Figure 2A**)., Recently, gut-brain cross-talk has been increasingly suggested by a growing number of studies 20. Our study showed that DMRs between LGA and HGA samples were significantly enriched in the dopaminergic synapse and serotonergic synapse pathways, which play a role in gut-brain axis model. HGA *vs.* Normal includes almost genes the LGA VS Normal and LGA VS HGA DMRs located on (**Figure 1G**). To figure out potential functional changes from low-grade adenoma to high-grade adenoma, the Gene Ontology (GO) enrichment were performed for 275 genes significantly differentially methylated just in LGA vs. Normal and HGA vs. Normal without LGA vs. HGA, and 571 significantly differentially methylated genes shown in LGA vs. HGA and HGA vs. Normal without LGA vs. Normal (**Figure 2B**). For 275 genes significantly differentially methylated just in LGA vs. Normal and HGA vs. Normal, GO enrichment analysis showed that highest ranked GO terms included proteolysis, extracellular matrix disassembly, inorganic anion transport and cobalamin metabolic processes. . Cell adhesion, and positive regulation of positive chemotaxis and neuropeptide signaling pathway were also encompassed on the overlapping region between LGA vs. Normal and LGA vs. HGA. Notably, results showed that genes significantly differentially methylated only from low-grade adenoma to high-grade adenoma were enriched for chemical synaptic transmission, transmission of nerve impulse, calcium ion transmembrane transport and etc., most of which are related to nervous system function, and exhibited different pattern of LGA vs. HGA compared with LGA vs. Normal.

**Hyper-methylated CpG sites showed better diagnostic performance than the hypo-methylated pattern**

In order to evaluate the discriminatory ability of DNA methylation to distinguish between patterns observed in normal tissue, adenoma and colorectal cancer, we collected 833 genome-wide DNA methylation dataset from GEO and ArrayExpress, including 278 normal tissue samples, 51 adenoma samples and 504 cancer samples. We separated DMSs into two groups including hyper-DMSs and hypo-DMSs. We found both hyper-DMSs and hypo-DMSs exhibited capacity to effectively differentiate pathological specimens (adenoma and cancer) from normal specimens (**Figure 3A** and **Figure 3B**). Meanwhile, we conducted two machine learning based predictive modeling ofDMSs identified in our dataset as predictive variables. Outcomes of predictive modeling applying random forest and neural network approaches revealed that hyper-methylated sites had superior capacity in distinguishing between normal samples and the disease samples. (**Table 1**). Area under the curve (AUC) of receiver operating characteristic (ROC) curve for hyper-methylated sites applying random forest and neural network were 0.91 and 0.85, respectively, whereas for hypo-methylated sites, AUC of ROC curve just were 0.72 and 0.76, respectively (**Figure 3C** and **Figure 3D**). Unsupervised tSNE cluster analysis mirrored the same patterns (**Figure 3E** and **Figure 3F**). To avoid inconsistent results caused by unstable methylation based on single CpG sites, we compared mean beta value (mBV) of these sites. We found that the hyper-methylated mBVs exhibited a statistically significant difference between normal tissue and cancers (P<2.2x10-16). However no differences between methylation patterns were not observed in on comparison of those noted in the context of adenoma and cancer (P= 0.29, **Figure 3G**) in which the average mBV of the normal tissue, the adenoma and the cancer are 0.22, 0.54 and 0.57 respectively. We observed similar results for hypo-methylation sites in which the average mBV of the normal tissue, the adenoma and the cancer were 0.70, 0.44 and 0.50 respectively (**Figure 3G**). Finally, we found the AUC of ROC curve with hyper-mBV and hypo-mBV were 0.98 and 0.95, respectively. Permutation analysis based on bootstrap strategy showed the model based on hyper-methylated sites had better discrimination than the model of hypo-methylated sites (P<2.2x10-8, **Figure 3H**).

#### The promoter of ADHFE1 maybe a potential biomarker for colorectal adenoma and cancer

Next, we grouped DMRs between the normal tissue and the LGA into hyper and hypo DMRs, and performed enrichment analysis by Ingenuity Pathway Analysis, respectively (Method). The top enriched functional term for hyper DMRs was ethanol degradation II (P=5.4x10-3), mostly contributed by two genes, *ADHFE1* and *ACSS3*, which can facilitate translation from(?) ethanol to ethanol and from acetic acid to acetyl-CoA respectively (**Figure 4A**). Both of them showed expression down regulation in colonic and rectal cancer tissue compared with the normal tissue (P<0.01), which are consistent with the DNA methylation changes (R2=0.49 and -0.59, **Figure 4B** and **Figure 4C**). We found the average methylation level of CpG sites located in CpG islands within *ADHFE1* and*ACSS3* promoter region were significantly increased in cancer samples compared with normal samples (mBVs=0.2 and 0.18 respectively). Furthermore, we applied promoter region within CpG island of the two genes to distinguish the normal tissue and the disease tissues (adenoma and cancer). When setting a cutoff at 0.25 for *ADHFE1* promoter, the minimal error rate was only 4.68% (39/833, **Figure 4D**). Theheatmap of sites within the region reflected similar results (**Figure 4E**). Furthermore, analysis of the ROC curve of mBV of *ADHFE1* promoter of all 833 samples resulted in an AUC of 0.97, achieving specificity and sensitivity of 0.95 and 0.96, respectively (**Figure 4F**). Within cancer samples, AUC for mBV of *ADHFE1* promoter achieved 0.98 (**Figure S2**). For *ACSS3*, the minimal error rate of its promoter was 16.68% (139/833) with the cutoff at 0.42 (**Figure 4G**), showing inferior performance compared to discriminatory power to  *ADHFE1* . Meanwhile, we also compared *ADHFE1* with *SEPT9,* which has been approved by FDA as a methylation-based CRC screening biomarker. We found ADHFE1 had better predictive power than *SEPT9* (**Figure 4H**) 21. Furthermore, we observed *ADHFE1* exhibited a much better separation boundary compared with *SEPT9* (**Figure 4I**).

## **Discussion**

Whole genome DNA hypomethylation and hypermethylation of promoter of cancer-related gene are regarded as the common pattern across diverse cancers. In our study, we found whole genome DNA hypomethylation may start at benign adenoma stage and high-grade adenoma showed increased hypomethylation compared to low-grade adenoma (**Figure 1C**). As many previous studies reported, bimodal distribution can characterize DNA methylation pattern, and we found (that decline in hypermethylated peak can clearly reflect progressive hypomethylation (**Figure 1D** and **Figure 1E**)22. We identified 440 and 6,805 DMRs in low- and hyper-grade adenoma respectively, and 314(71.4%) in low-grade adenoma and 4,213(61.9%) in high-grade adenoma were hypomethylated. By contrast, most of DMRs (660/868, 76.0%) in high-grade adenoma and low-grade adenoma were hypermethylated. In addition, there was a little overlap between the genes and? the significant distinct DMRs located on LGA *vs.* Normal and LGA vs. HGA. Both of these results indicate LGA *vs.* Normal and LGA *vs.* HGA possibly are not the same process with degree difference but two different epigenetic processes. To find functional differences, enrichment analysis was applied for 603 genes with DMRs between high-grade adenoma and low-grade adenoma, and most enriched terms related to nervous system and signal transduction (**Figure 2A**). The “gut–brain-axis” conceptualizes an integrated physiological system encompassing afferent and efferent neural, endocrine, nutrient, and immunological signaling between the CNS and GI tract 20. Our KEGG enrichment analysis showed the significance of dopaminergic synapse and serotonergic synapse, and both are important in nervous system. Serotonin (5-hydroxytryptamine, 5-HT) has a popular image as a contributor to feelings of well-being and happiness, although its actual biological function is complex and multifaceted, and includes modulation of cognition, reward determination, learning, memory, and numerous additional physiological processes23. The role of Brain 5-HT, has been more widely studied and publicized despite the comparatively larger store of 5-HT in the gut24. Dopamine (3,4-dihydroxyphenethylamine, DA) is an organic chemical of the catecholamine and phenethylamine families. It functions both as a hormone and a neurotransmitter, and plays several important roles in the brain and body. In the brain, dopamine functions as a neurotransmitter, a chemical released by neurons (nerve cells), to send signals to other nerve cells. Outside the central nervous system, dopamine functions primarily as a local paracrine messenger25. It reduces gastrointestinal motility and protects intestinal mucosa. Findings of nervous system-associated terms in our study were unexpected and will require further study to define the concrete mechanism. Our study suggests that the gut–brain-axis and molecules related to this interaction may represent a new target to explore with respect to early diagnosis and risk assessment for colorectal cancer, even at benign adenoma stage.

DNA methylation has always been considered as a potential biomarker for many diseases due to its tissue specificity and status stability. Here, we used it to distinguish disease samples (including adenoma and cancer) from normal samples. We compared 209 hyper-methylated sites and 441 hypo-methylated sites between LGA and normal samples, and found both hyper-methylated sites and hypo-methylated sites with capacity to effectively differentiate between normal and LGA samples . Subsequently, we applied random forest and neural network analysis to verify our observation.Notably, AUCs of ROC curves for prediction model applied to hyper-methylated sites exceeded those applied to hypo-methylated sites, despite the observation of a >two-fold frequency of hypo-methylated sites compared to hyper-methylated sites. A growing evidence base supports whole-genome hypo-methylation as a common characteristic shared by tumors . We speculate that hyper-methylation at several key sites or genes observed at early colorectal adenoma may be pivotal early events, and hypo-methylation may be more widely-associated incidental events of disease progression.

To preclude the bias caused by unstable methylation on single CpG site, we compared mean beta value (mBV) of these sites. We found that levels of the hyper-methylated mBVs were significantly different between normal tissue and cancers (P<2.2x10-16), while no significant difference was found between the adenoma and the cancer (P= 0.288, **Figure 3G**). The average mBV of the normal tissue, the adenoma and the cancer are 0.218, 0.542 and 0.568 respectively. Similar results were observed for hypo-methylation sites, where the average mBV of the normal tissue, the adenoma and the cancer were 0.698, 0.444 and 0.499 respectively (**Figure 3G**). Finally, we found the AUCs of ROC curve with hyper-mBV and hypo-mBV were 0.982 and 0.947, respectively. Permutation analysis based on bootstrap strategy found that the model based on hyper-methylated sites showed better discriminatory capacity than the model of hypo-methylated sites (P<2.2x10-8, **Figure 3H**).

Colorectal adenoma is believed to be the intermediate stage between normal status and cancer. Our study focused on differences in DNA methylation patterns observed in adenoma stages and across the spectrum of more advanced cancer stages. Ethanol degradation II was the top term observed following IPA enrichment of hyper-DMRs identified in very early stage, and *ADHFE1* and *ACSS3* genes were dominantly featured in this biological process. Intense early DNA methylation changes at these loci provides observation of these changes as a potential adenoma biomarker. After observing negative correlation of expression and DNA methylation of the two genes, we only used their methylation levels to distinguish normal and adenoma samples. The error rate of *ADHFE1* was 4.68% (39/833), and’s 16.68% (139/833) for *ACSS3*. Furthermore, according to ROC curve of mBV of *ADHFE1* promoter mBV for all 833 samples, the AUC was 0.968 with specificity and sensitivity as 0.946 and 0.960 (**Figure 4F**). Performance for distinguishing between cancer and normal samples was even better, and the AUC was 0.978 (**Figure S3**).

The *ADHFE1* gene is known to encode for hydroxyacid-oxoacid transhydrogenase, responsible for the oxidation of 4-hydroxybutyrate in mammalian tissues. Association of this gene with cellular proliferation and differentiation has been reported 26. In colorectal cancer tissue, *ADHFE1* gene shows hyper-methylation in promoter region and down regulation of expression, and it may facilitate tumor growth 27. Our results suggest the methylation level of *ADHFE1* promoter may qualify this finding as a a potential biomarker for distinguishing adenoma from normal tissue in either traditional biopsy or liquid biopsy. Validation in large sample size will be needed to access its efficacy.

## **Methods**

**Sample collection and pathological confirmation**

We collected 20 normal tissue specimens, 18 LGA specimens and 22 HGA specimens from patients who underwent endoscopic treatment in the Department of Gastroenterology of Peking University Third hospital from March 2015 to June 2016. Tissue specimens were embedded in paraffin, sectioned, stained with hematoxylin and eosin, and adenoma status was confirmed by pathologists following light microscopic evaluation. Sample information and loading quantity are provided on **Table S4**. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Ethics Committee of Peking University Third hospital (IRB number 206H005). Written informed consent was obtained from all patients and volunteers.

**DNA isolation and bisulfite conversion**

DNA was isolated using QIAmp DNA Mini Kit (you need to name the manufacturer, city), according to manufacturer’s protocol. Bisulfite conversion was performed using the EZ DNA Methylation-Gold Kit (Manufacturer, city) according to the instruction manual.

**Methylation data processing**

Epigenome-wide DNA methylation assessment for this study was performed using the Illumina Infinium Human Methylation 450 BeadChip (Illumina, San Diego, CA, USA), which simultaneously profiles the methylation status for >485,000 CpG sites at single-nucleotide resolution, covering 96% of CpG islands, with additional coverage of island shores (<2 Kb from CpG Islands), island shelves (2–4 Kb from CpG islands), and regions flanking them. The methylation status for each CpG site was calculated as the ratio of fluorescent signals (β = Max(M,0)/[Max(M,0) + Max(U,0) + 100]), ranging from 0 to 1, using the average probe intensity for the methylated (M) and unmethylated (U) alleles. β= 1 indicates complete methylation; β = 0 represents no methylation. The raw data from the array was processed using The GenomeStudio Methylation module (manufacturer, city). Methylation level calculations, normalization and background adjust was performed by the software. Probes located on sex chromosomes or those that failed detection P value testing for at least 1 sample or those representing a single nucleotide polymorphism (SNP), were removed from the analysis using R package IMA (vision 3.1.2) (manufacturer, city)28. DMRs were defined as rank sum test following false discovery rate (FDR)-adjusted P value<0.05 and |β|>0.15, and DMSs were defined as rank sum test following FDR adjust P value<0.05 and |β|>0.20. Promoter regions were defined as 5’UTR, TSS200, TSS1500 and first exons.

**Public data collection and processing**

In order to ensure consistency of data processing, we only collect sample with raw idat files, and then GSE68060, GSE68838, GSE77954, GSE77965, GSE81211, GSE101764, GSE107352 and GSE75546 were collected from GEO, E-MTAB-6450 was collected from ArrayExpress (Archive of Functional Genomics Data) 29, 30, 31, 32, 33, 34. The information extracted from these public data was provided in Supplementary Table 5. Some cell line samples and metastatic cancer samples in above datasets were removed after further qanalysis. We collected data on a total of 278 normal samples, 51 adenoma samples and 504 cancer samples. All of these datasets obtained by accessing raw *idat* files were preprocessed using R package minfi (version 1.28.4) 35. The sites which failed detection (P = 0.01) were rewritten by nearest neighbor average to ensure an adequate number of sites.

**Comparation of the discriminatory ability**

For Random Forest prediction, we use R package RandomForest (version 4.6.14) and Number of trees are 500036. For neural network prediction, we use R package nnet (vision 7.3.12) with number of units in the hidden layer as 2 and weight decay as 10-4 and maximum number of iterations as 40037. The R package pROC (version 1.14.0) was used to do Receiver operating characteristic curve (ROC) analysis to compare the abilities between hyper and hypo- sites by AUC38.

**t-Distributed Stochastic Neighbor Enbedding (t-SNE) analysis, (Principal component analysis (PCA) and Gene Enrichment analysis**

t-SNE analysis was performed by R package t-SNE (vision 0.1-3)39. PCA was performed by R function princomp and visualized by first two principal components. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene ontology (GO) enrichment were analyzed online by DAVID 6.8 (<https://david.ncifcrf.gov>)40, 41. Ingenuity Pathway Analysis (IPA) was also used for enrichment analysis for more elaborate result, setting the cutoff of P value as 0.0542.

## Abbreviation Table:

LGA: Low-grade adenoma

HGA: High-grade adenoma

LGA VS Normal: Comparison of low-grade adenoma with normal tissue

HGA VS Normal: Comparison of high-grade adenoma with normal tissue

LGA VS HGA: Comparison of high-grade adenoma with low-grade adenoma

DMR: Different methylation region

DMS: Different methylation site

ROC: Receiver operating characteristic

AUC: Area under the curve

IPA: Ingenuity Pathway Analysis

KEGG: Kyoto Encyclopedia of Genes and Genomes

GO: Gene Ontology

t-SNE: t-distributed stochastic neighbor embedding

PCA: Principal components analysis

mBV: Mean beta values

## Data and Code Available

**Data Access**

DNA methylation data and the scripts are available upon request. Other data involved in this study included GSE68060, GSE68838, GSE77954, GSE77965, GSE81211, GSE101764, GSE107352, GSE75546 and E-MTAB-6450.

## **Author Contribution**

JF developed analysis method and drafted the manuscript. JL recruited patients and provided clinical expertise on colorectal adenoma and interpreted the results. JF, HZ, and SG performed data analysis. SG and DZ reviewed and edited the manuscript. CT and YZ conducted array experiments. WW collected and prepared tissue samples and collected results of clinical assays. SD enrolled patients and collected all the clinical information. DZ and CZ designed the study, supervised all experiments and analysis, providing molecular and cellular biology advice, reviewed and edited the manuscript.

## **Acknowledgement**

This study is funded by Innovation Promotion Association CAS (2016098) , Major State Basic Research Development Program (2014CB542006), Key Research Program of the Chinese Academy of Sciences (KJZD-EW-L14), and National Key Research and Development Plan of China (2016YFA0201404).

## **References**

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7-30 (2018).

2. Chen W*, et al.* Cancer statistics in China, 2015. *CA Cancer J Clin* **66**, 115-132 (2016).

3. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nature genetics* **49**, 635-642 (2017).

4. Wang X*, et al.* Hypermethylation reduces expression of tumor-suppressor PLZF and regulates proliferation and apoptosis in non-small-cell lung cancers. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **27**, 4194-4203 (2013).

5. Guo S*, et al.* Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC). *Clinical epigenetics* **7**, 3 (2015).

6. Zhao Y*, et al.* Genome-wide methylation profiling of the different stages of hepatitis B virus-related hepatocellular carcinoma development in plasma cell-free DNA reveals potential biomarkers for early detection and high-risk monitoring of hepatocellular carcinoma. *Clinical epigenetics* **6**, 30 (2014).

7. Haikun Zhang PD, Shicheng Guo, Chengcheng Tao, Wenmin Zhao, Jiakang Wang, Ramsey Cheung, Augusto Vilanueva, Huiguo Ding, Steven J. Schrodi, Dake Zhang, Changqing Zeng. Circulating cell-free DNA based low-pass genome-wide bisulfite sequencing aids non-invasive surveillance to Hepatocellular carcinoma. *Science Advance (Submitted)*, (2019).

8. Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* **135**, 1079-1099 (2008).

9. Hidaka H*, et al.* Comprehensive methylation analysis of imprinting-associated differentially methylated regions in colorectal cancer. *Clin Epigenetics* **10**, 150-150 (2018).

10. Shi YX*, et al.* Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer. *BMC Genomics* **18**, 901 (2017).

11. Lindqvist BM, Wingren S, Motlagh PB, Nilsson TK. Whole genome DNA methylation signature of HER2-positive breast cancer. *Epigenetics* **9**, 1149-1162 (2014).

12. Raggi C, Invernizzi P. Methylation and liver cancer. *Clin Res Hepatol Gastroenterol* **37**, 564-571 (2013).

13. Morris MR, Latif F. The epigenetic landscape of renal cancer. *Nat Rev Nephrol* **13**, 47-60 (2017).

14. Herman JG*, et al.* Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Research* **55**, 4525 (1995).

15. Kane MF*, et al.* Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Research* **57**, 808 (1997).

16. Yoshiura K, Kanai Y, Ochiai A, Shimoyama Y, Sugimura T, Hirohashi S. Silencing of the E-cadherin invasion-suppressor gene by CpG methylation in human carcinomas. *Proceedings of the National Academy of Sciences* **92**, 7416 (1995).

17. Witold K, Anna K, Maciej T, Jakub J. Adenomas - Genetic factors in colorectal cancer prevention. *Rep Pract Oncol Radiother* **23**, 75-83 (2018).

18. Rex DK*, et al.* American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am J Gastroenterol* **104**, 739-750 (2009).

19. Perez-Silva JG, Araujo-Voces M, Quesada V. nVenn: generalized, quasi-proportional Venn and Euler diagrams. *Bioinformatics* **34**, 2322-2324 (2018).

20. Clemmensen C, Muller TD, Woods SC, Berthoud HR, Seeley RJ, Tschop MH. Gut-Brain Cross-Talk in Metabolic Control. *Cell* **168**, 758-774 (2017).

21. Church TR*, et al.* Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut* **63**, 317-325 (2014).

22. Straussman R*, et al.* Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol* **16**, 564-571 (2009).

23. Swami T, Weber HC. Updates on the biology of serotonin and tryptophan hydroxylase. *Curr Opin Endocrinol Diabetes Obes* **25**, 12-21 (2018).

24. Xiaolong G, Junhai P, Yichang L, Hongkan W, Wei Z, Xianfa W. Intestinal Crosstalk between Microbiota and Serotonin and its Impact on Gut Motility. *Current Pharmaceutical Biotechnology* **19**, 190-195 (2018).

25. Berke JD. What does dopamine mean? *Nat Neurosci* **21**, 787-793 (2018).

26. Deng Y*, et al.* Cloning and characterization of a novel human alcohol dehydrogenase gene (ADHFe1). *DNA Seq* **13**, 301-306 (2002).

27. Tae CH*, et al.* Alcohol dehydrogenase, iron containing, 1 promoter hypermethylation associated with colorectal cancer differentiation. *BMC Cancer* **13**, 142 (2013).

28. Wang D*, et al.* IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28**, 729-730 (2012).

29. Qu X*, et al.* Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. *Oncogene* **35**, 6403-6415 (2016).

30. consortium B. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat Biotechnol* **34**, 726-737 (2016).

31. Kang K, Bae JH, Han K, Kim ES, Kim TO, Yi JM. A Genome-Wide Methylation Approach Identifies a New Hypermethylated Gene Panel in Ulcerative Colitis. *Int J Mol Sci* **17**, (2016).

32. Barrow TM*, et al.* Smoking is associated with hypermethylation of the APC 1A promoter in colorectal cancer: the ColoCare Study. *J Pathol* **243**, 366-375 (2017).

33. Damaso E*, et al.* Primary constitutional MLH1 epimutations: a focal epigenetic event. *Br J Cancer* **119**, 978-987 (2018).

34. Bormann F*, et al.* Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep* **23**, 3407-3418 (2018).

35. Aryee MJ*, et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).

36. Wiener ALaM. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).

37. Ripley WNVaBD. *Modern Applied Statistics with S*, Fourth edn. Springer (2002).

38. Robin X*, et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

39. Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

40. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

41. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).

42. Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523-530 (2014).

## **Figure and legends:**

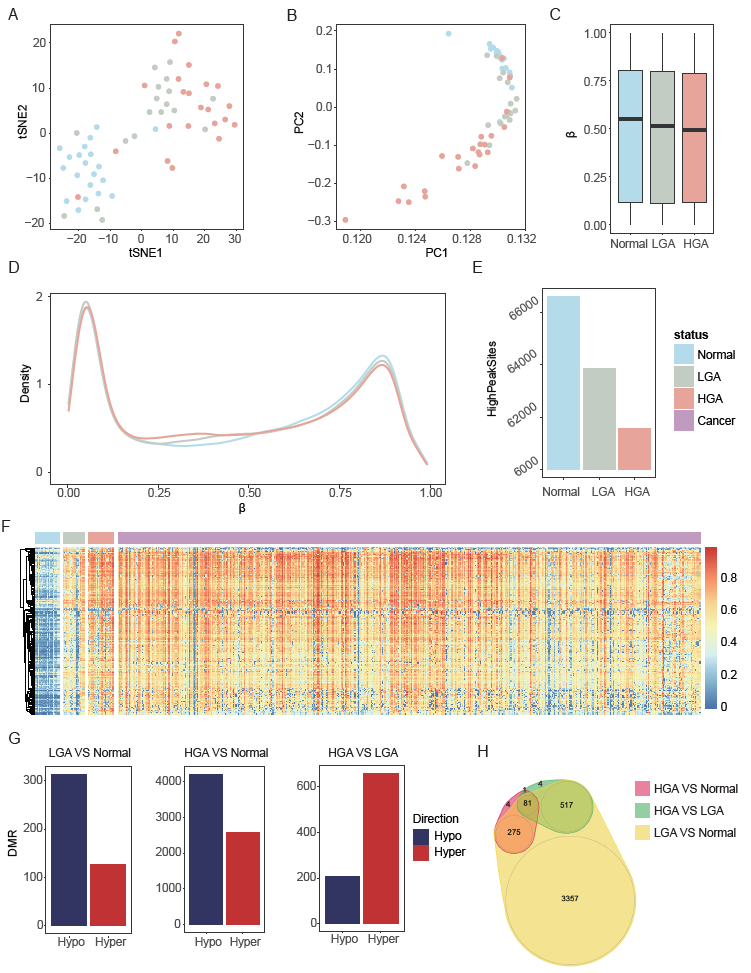


Figure 1. Genome-wide DNA methylation of low-grade adenoma (LGA), high-grade colorectal adenoma (HGA) and normal colorectal tissue. (A): tSNE analysis to show the data structure and sample relationship. (B): PCA analysis to show the data structure and sample relationship. (C): Average methylation level of N, LGA and HGA. (D): Density plot to show the distribution of the whole array probes cross N, LGA and HGA. (E): Number of sites in β ranging from 0.7 to 0.9. (F): The heatmap of the 209 hyper-methylated DMSs of in-house datasets 504 public cancer datasets. (G): DMR between LGA and normal tissues, HGA and normal tissue, and HGA and LGA. (H): Venn graph to show all the above DMRs.

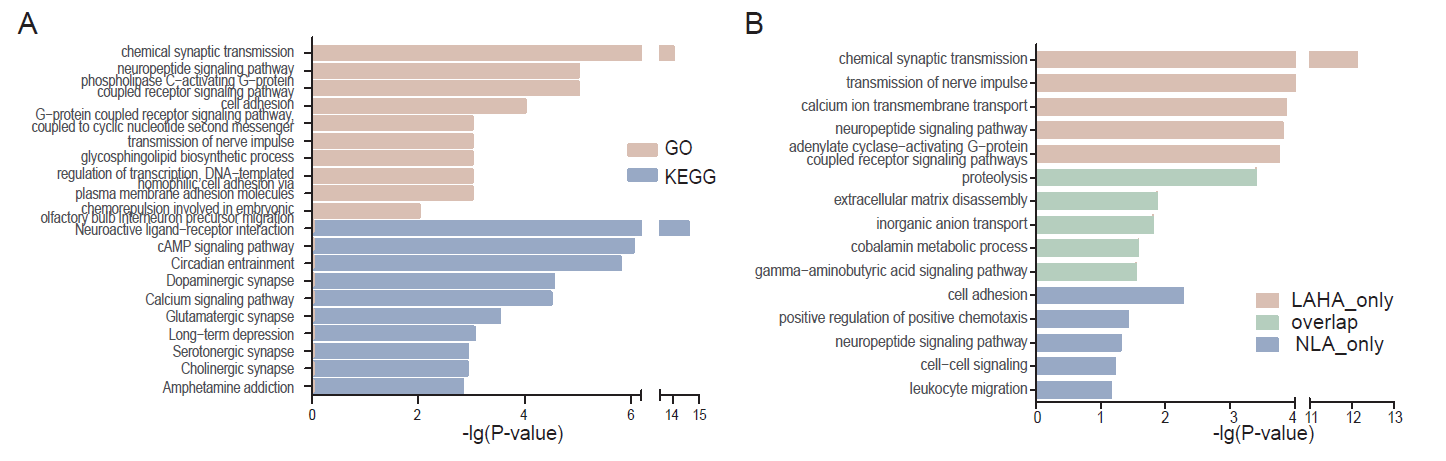


Figure 2. Enrichment analysis shown Nervous system was associated with adenoma development. (A) GO and KEGG analysis of the genes the LGA VS HGA DMRs located on. (B) GO analysis of the genes different DMRs located on, including the DMR only in LGA VS HGA, only in LGA VS Normal, and LGA VS HGA and LGA VS Normal overlapped.

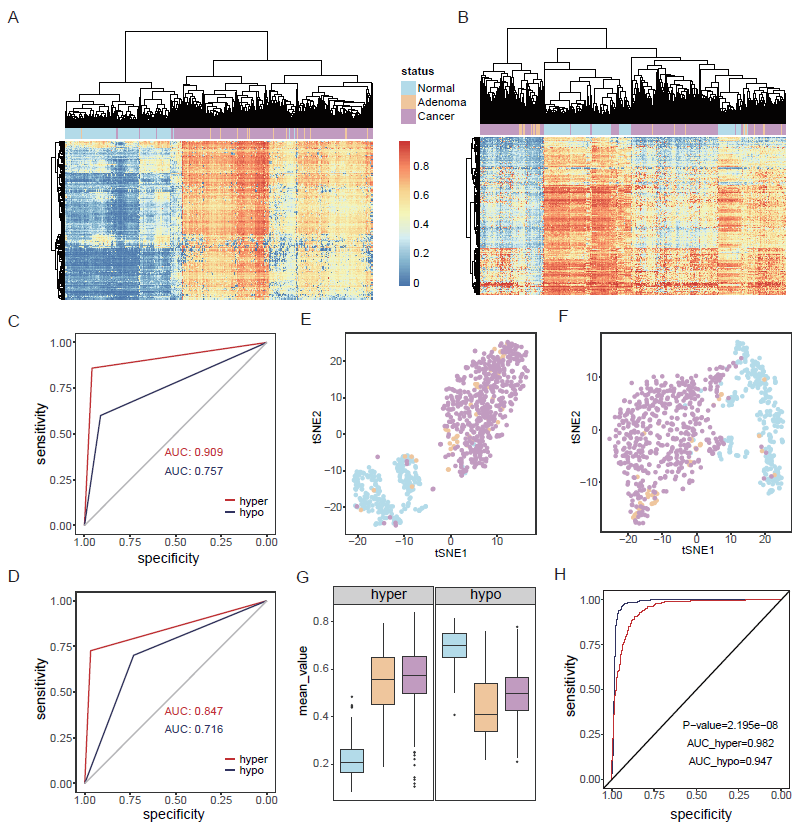


Figure 3. Hyper-methylated CpG sites showed better diagnostic performance than the hypo-methylated pattern. (A): Cluster analysis based on hyper-DMSs among normal, adenoma and cancer samples. (B): Cluster analysis based on hypo-DMSs among normal, adenoma and cancer samples. (C): Random forest prediction performance based on hyper and hypo-DMSs. (D): Neural network prediction performance based on hyper and hypo-DMSs. (E): tSNE analysis to show the data structure and sample relationship based on hyper-DMSs. (F): tSNE analysis to show the data structure and sample relationship based on hypo-DMSs. (G): Average methylation level of hyper and hypo-DMSs (H): ROC curve of hyper-mBV and hypo-mBV.

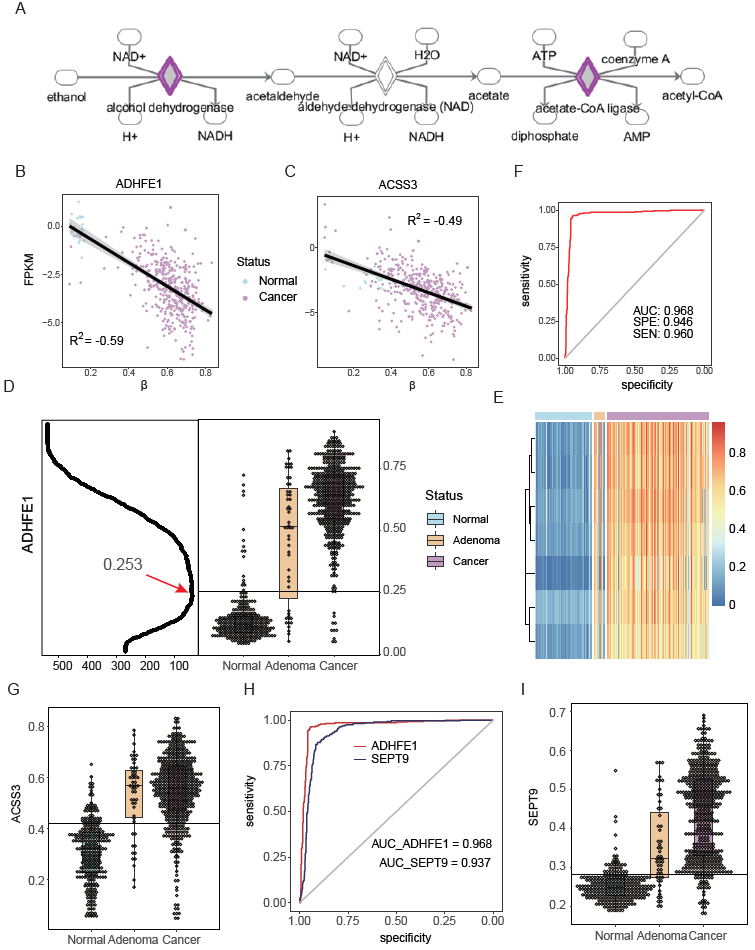


Figure 4. DNA methylation *ADHFE1* and *ACSS3* in Normal, LGA and HGA. (A): pathway of ethanol degradation II. (B): relationship between DNA methylation and gene expression of *ADHFE1*. (C): relationship between DNA methylation and gene expression of *ACSS3*. (D): Left part is identification of cutoff, X axis is sample number of classification error; Right part is DNA methylation of *ADHFE1* in normal, adenoma and cancer samples. (E): The heatmap of sites within ADHFE1 promoter in normal, adenoma and cancer samples. (F): ROC of the prediction of *ADHFE1* for colorectal adenoma and caner. (G): DNA methylation of *ACSS3* in normal, adenoma and cancer samples. (H): ROC comparison of *ADHFE1* and *SEPT9*. (I): DNA methylation of *SEPT9* in normal, adenoma and cancer samples.

Table 1. Prediction performance based on hyper-DMS and hypo-MDS to distinguish disease and normal

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Methylation | Observation | Prediction | | Sensitivity | Specificity |
| Disease | Normal |
| Random Forest | hyper | Disease | 532 | 23 | 0.959 | 0.860 |
| Normal | 39 | 239 |
| hypo | Disease | 507 | 48 | 0.914 | 0.601 |
| Normal | 111 | 167 |
| Neural Network | hyper | Disease | 537 | 18 | 0.968 | 0.727 |
| Normal | 76 | 202 |
| hypo | Disease | 406 | 149 | 0.732 | 0.701 |
| Normal | 83 | 195 |