**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 Guo3,4 # ,Ingrid Glurich3, Chengcheng Tao1, Haikun Zhang1,5, Wenjing Wang2, Ying Zhang1, Steven J Schrodi3,4,6, 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，Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA

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

6，Department of Medical Genetics, School of Medicine and Public Health, University of Wisconsin-Madison, WI 53706, 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 abnormal is the hallmark of human cancers and was demonstrated to be most promising biomarker for early diagnosis to human cancers. However, majority of DNA methylation biomarkers were identified by the hypothesis that early differential methylation regions (DMRs) are maintained and could be detected in all stage of 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 are progressively accumulated in tumor genome during the occurrence of human cancers. DNA methylation plays important roles in embryonic development and tissue differentiation. Abnormal, hyper-methylation or hypo-methylation, in the promoter regions of tumor suppressor genes and miRNA have been observed in almost all the cancer types [3, 4]. In the past decades, DNA methylation has been widely applied to develop cancer biomarkers [5]. Meanwhile, it has perfect ability to indicate disease progression, such as from hepatitis, cirrhosis and HCC [6, 7]. Moreover, recent evidence shows 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. Global hypo-methylation arises early, which 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) [10-12]. 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]. At the background of whole genome hypo-methylation, gene-specific promoter hyper-methylation has been found to promote CRC by down-regulation the expression of key tumor suppressor gene, such as *CDKN2A*, *MLH1*, and *CDH1*[14-16]. Although extensive epigenetic alterations have been illustrated over the past years, CRC is still not well understood at the molecular level. CRC is a heterogeneous disease, which typically starts from the [benign tumor](https://en.wikipedia.org/wiki/Adenoma), often in the form of the adenoma, and past more than 10 years becomes malignant cancer [17]. Even CRC in both incidence and mortality are higher in all kinds of cancer, adenoma stage provides an excellent opportunity to prevent its cancerization and get excellent survival. A large of studies were focusing on CRC, while a part of them treated adenoma as middle stage lacking of further specific study. Actually, colorectal adenoma has different pathologic stages (low-grade adenoma and high-grade adenoma), and no research has compared the different adenomas of different stages at Whole-genome DNA methylation level[18]. Besides, alterations on low-grade adenoma maybe potential diagnostic biomarker. Therefore, the comprehensive understanding to the genome-wide DNA methylation profile for colorectal cancer, especially the early stage pre-cancerous lesions (low-grade adenoma and high-grade adenoma), will provide important resources for cancer early diagnosis and candidate biomarkers for cell-free DNA methylation research.

In this study, we firstly treated adenoma as two stages, and conducted genome-wide DNA methylation array to 18 low-grade adenoma (LGA) and 22 high-grade colorectal adenoma (HGA) and 20 normal tissue from Chinese population. Dynamic DNA methylation change of colorectal low and high-grade adenoma was identified and we found the methylation change appeared in LGA will strengthen or maintain in LGA and caner. We conducted enrichment analysis to differential methylation regions (DMRs) to inquiry potential DNA methylation influenced functional difference in adenoma initiation and development stages. Moreover, we separated different methylation sites (DMSs) in LGA VS Normal into hyper-DMS and hypo-DMS, and evaluated their respective performance for the colorectal adenoma and cancer 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 on single-base level for 18 low-grade adenoma (LGA) and 22 high-grade colorectal adenoma (HGA) and 20 normal tissue (**Method**). We found the significant genome-wide DNA methylation difference between normal, low and high-grade adenoma (**Figure 1A** and **1B**). Compared with the normal tissue, low-grade adenoma had genome wide hypo-methylation (P = 2.8x10-51), and global methylation level become even lower in high-grade adenoma (P = 1.6x10-88, compared with normal, t-test 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 malignancy degree of lesions (right peak, **Figure1D**; **Figure 1E**). Almost all DMSs in LGA compared with normal strengthen or maintain in HGA and cancer, highlighting the importance of the in-depth exploration on pre-cancerous benign lesion. The 209 significantly hyper-methylated sites in LGA showed further methylated in LGA and 504 cancer samples collected from the public databases (**Figure 1F**) and hypo-DMSs showed diametric tendency (**Supplementary Figure 2**). It well showed that DNA demethylation may occur very early in precancerous lesions. Particularly, 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, Supplementary Table 1-2**). However, with LGA as the reference, most DMRs observed in HGA were hyper-methylated (76.0%, 660/868) (**Figure 1G, Supplementary Table 3**). This suggested genome wide demethylation may dominate the canceration course of tissue cells, but hyper-methylation sites may contribute more to the distinct malignancy of these lesions. In addition, there were limited overlapping between DMRs in LGA compared to normal tissues and those compared to HGA. hinting the different epigenetic processes (**Figure 1H**)[19].

#### Nervous system is associated with adenoma development

Enrichment analysis to 603 DMRs between HGA and LGA located on, most terms are nervous system and signal transduction associated (**Figure 2A**). Recent years, gut-brain cross-talk is focused by more and more studies [20], and in our study dopaminergic synapse and serotonergic synapse are hit on KEGG enrichment result, 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 function changes from low-grade adenoma to high-grade adenoma, the Gene Ontology (GO) enrichment were performed for 275 genes significantly different methylated just in LGA VS Normal and HGA VS Normal without LGA VS HGA, and 571 significantly different methylated genes shown in LGA VS HGA and HGA VS Normal without LGA VS Normal (**Figure 2B**). For 275 genes significantly different methylated just in LGA VS Normal and HGA VS Normal, GO analysis shows the top term enriched is proteolysis, and extracellular matrix disassembly, inorganic anion transport and cobalamin metabolic process also be hit. Cell adhesion, and positive regulation of positive chemotaxis and neuropeptide signaling pathway are hit on overlapped part between LGA VS Normal and LGA VS HGA. What is intriguing is the results show the genes significantly different 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 them are nervous system associated, exhibiting 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 evaluation the distinguish ability of DNA methylation for 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 separate DMSs into two groups including hyper-DMSs and hypo-DMSs. We found both hyper-DMSs and hypo-DMSs could provide effective distinguish ability between disease samples (adenoma and cancer) and normal samples (**Figure 3A** and **Figure 3B**). Meanwhile, we conducted two machine learning based predictions with DMSs identified in our dataset, and we observed hyper-methylated sites can provide better distinguish between normal samples and the disease samples in prediction of random forest and neural network (**Table 1**). We found, for hyper-methylated sites the area under the curve (AUC) of receiver operating characteristic (ROC) curve are 0.91 and 0.85, respectively. For hypo-methylated sites, AUC of ROC curve just are 0.72 and 0.76, respectively (**Figure 3C** and **Figure 3D**). Unsupervised tSNE cluster analysis visually show the same result (**Figure 3E** and **Figure 3F**). In order to avoid the inconsistent result caused by unstable methylation based on single CpG site, we compared mean beta value (mBV) of these sites. We found that the hyper-methylated mBVs were significant different between normal tissue and cancers (P<2.2x10-16) while no significance was found between the adenoma and the caner (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 are 0.70, 0.44 and 0.50 respectively (**Figure 3G**). Finally, we found the AUC of ROC curve with hyper-mBV and hypo-mBV are 0.98 and 0.95, respectively. Permutation analysis based on bootstrap strategy shown the model based on hyper-methylated sites has 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 separate DMRs between the normal tissue and the low-grade adenoma into hyper and hypo DMRs. The enrichment analysis was performed by Ingenuity Pathway Analysis (IPA) for different DMRs, setting the cutoff of P value as 0.05. The first term of the IPA enrichment result for hyper DMRs is ethanol degradation II (P=5.4x10-3), where two genes are hit, *ADHFE1* and *ACSS3*, which can facilitate translation form ethanol to ethanal and from acetic acid to acetyl-CoA respectively (**Figure 4A**). Both of them showing expression down regulation on 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*, *ACSS3* promoter region are significantly increased in cancer samples compared with normal samples (mBVs=0.2 and 0.18 respectively, **Figure 4D**). 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 cutoff as 0.25 for *ADHFE1* promoter, we minimize the error rate to 4.68% (39/833), while the minimal error rate of *ACSS3* promoter is 16.68% (139/833) when setting cutoff as 0.42 (**Figure 4E**). The result shown *ADHFE1* has better discrimination power compared with *ACSS3*. Furthermore, at ROC curve of mBV of *ADHFE1* promoter mBV for all 833 samples, the AUC is 0.97 with specificity and sensitivity as 0.95 and 0.96 (**Figure 4F**). For cancer samples, it can reach even AUC as 0.98 (**Supplementary Figure 2**).

## **Discussion**

Whole genome DNA hypomethylation and hypermethylation of promoter of cancer related gene are regard as the common pattern of diverse cancers. In our study, we found whole genome DNA hypomethylation arising at benign adenoma stage and high-grade adenoma shows further hypomethylation compared to low-grade adenoma (**Figure 1C**). As many previous studies reported, bimodal distribution can characterize DNA methylation pattern, and we found hypermethylated peak can clearly reflect progressive hypomethylation (**Figure 1D** and **Figure 1E**)[21]. 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 are hypomethylated. Besides, we found 868 DMRs when compared high-grade adenoma with low-grade adenoma. What is interesting is most of them, 660 (76.0%), are hypermethylated, which is converse with LGA VS Normal and HGA VS Normal. Beyond our expectation, there is a little overlap between the genes 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. It impels us to figure out the potential mechanism. So we did enrichment analysis for 603 genes the DMRs between high-grade adenoma and low-grade adenoma located on, and most terms are nervous system and signal transduction associated (**Figure 2A**). The term gut–brain-axis describes an integrative physiology concept that incorporates all, including afferent and efferent neural, endocrine, nutrient, and immunological signals between the CNS and the gastrointestinal system, which is focused by more and more studies[20]. In our study, dopaminergic synapse and serotonergic synapse are hit on KEGG enrichment result, both of them are important for nervous system. Serotonin (5-hydroxytryptamine, 5-HT) has a popular image as a contributor to feelings of well-being and happiness, though its actual biological function is complex and multifaceted, modulating cognition, reward, learning, memory, and numerous physiological processes[22]. Brain 5-HT gets much more respect, and certainly more press, than the vastly larger store of 5-HT in the gut[23]. 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 messenger[24]. It reduces gastrointestinal motility and protects intestinal mucosa. Nervous system associated terms are unexpected in our study, so we need further study to uncover the concrete mechanism. Our study suggests gut–brain-axis and related molecule maybe be the new thinking of early diagnosis and risk warning of colorectal cancer, even at benign adenoma stage.

DNA methylation always be considered as a potential biomarker for many diseases for its tissue specificity and status stability, at the same time harboring pathological sensibility, and we want to use it to distinguish disease samples (including adenoma and cancer) from normal samples. We filtrated 209 hyper-methylated sites and 441 hypo-methylated sites from LGA VS Normal, and we found both hyper-methylated sites and hypo-methylated sites could provide effective distinguish ability between normal samples and disease samples. We used random forest and neural network to verify our observation and AUCs of ROC curves of hyper-methylated are larger than the hypo-methylated in two machine-learning based prediction models. Although more than twice number than hyper-methylated sites, performance of hypo-methylated sites still is inferior than the hyper-methylated. Synthesizing whole-genome hypo-methylation, we speculate hypo-methylation may be the widely incidental events to several key sites or genes hyper-methylation at early colorectal adenoma. In order to avoid the inconsistent result caused by unstable methylation based on single CpG site, we compared mean beta value (mBV) of these sites. We found that the hyper-methylated mBVs were significant different between normal tissue and cancers (P<2.2x10-16) while no significance was found between the adenoma and the caner (P= 0.288, **Figure 3G**) in which the average mBV of the normal tissue, the adenoma and the cancer are 0.218, 0.542 and 0.568 respectively. We observed similar results for hypo-methylation sites in which the average mBV of the normal tissue, the adenoma and the cancer are 0.698, 0.444 and 0.499 respectively (**Figure 3G**). Finally, we found the AUC of ROC curve with hyper-mBV and hypo-mBV are 0.982 and 0.947, respectively. Permutation analysis based on bootstrap strategy shown the model based on hyper-methylated sites has better discrimination than the model of hypo-methylated sites (P<2.2x10-8, **Figure 3H**).

Most of colorectal adenoma consider adenoma as a middle stage between normal status and cancer. Our study focuses on adenoma and compare the different pathological stages. At very early stage, ethanol degradation II is the top term of IPA enrichment result of hyper-DMRs, in which *ADHFE1* and *ACSS3* are hit. The strong early DNA methylation change provides potential as biomarker of disease. After getting negative correlation of expression and DNA methylation of the two genes as expected, we try to verify the probability as to distinguish normal samples and disease samples. The error rate of *ADHFE1* just get 4.68% (39/833), while the *ACSS3*’s higher as 16.68% (139/833). Furthermore, at ROC curve of mBV of *ADHFE1* promoter mBV for all 833 samples, the AUC is 0.968 with specificity and sensitivity as 0.946 and 0.960 (**Figure 4F**). For cancer samples, it can reach even AUC as 0.978 (**Supplementary Figure 3**). The *ADHFE1* gene encodes hydroxyacid-oxoacid transhydrogenase, which is responsible for the oxidation of 4-hydroxybutyrate in mammalian tissues there are some studies report the gene is associated with cell proliferation and differentiation[25]. At colorectal cancer tissue, *ADHFE1* gene show hyper-methylated and down regulation of expression, by the way it may facilitated tumor growth [26]. Our results suggest the promoter of *ADHFE1* can be a potential biomarker, which can distinguish disease form normal tissue well, but adenoma sample is still limited and more sample should be taken in next study. One the other hand, liquid biopsy is the ideal way to apply *ADHFE1* to clinic scene, which is another direction of future research.

## **Methods**

**Sample collection and pathological confirmation**

We collected 20 normal tissue specimens, 18 low-grade adenoma specimens and 22 high-grade adenoma specimens from the 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 confirmed by pathologist by light microscopy. Sample information and loading quantity are provided on **Supplementary Table 4**. 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 xxxxx). Informed written consent were obtained from all patients and volunteers.

**DNA isolation and bisulfite conversion**

DNA was isolated using QIAmp DNA Mini Kit, according to manufacturer’s protocol. Bisulfite conversion was performed using the EZ DNA Methylation-Gold Kit 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, calculation of methylation levels, normalization and background adjust was performed by the software. Probes located on sex chromosomes or failed detection P value testing at least 1 sample or being SNP, were removed from the analysis using R package IMA (vision 3.1.2)[27]. DMRs were defined as rank sum test following FDR adjust 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 that 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 [28-33]. The information of these public data was provided on Supplementary Table 5. Some cell line samples and metastatic cancer samples in above datasets were removed at further study. All we collected 278 normal samples, 51 adenoma samples and 504 cancer samples. All of these datasets accessing raw *idat* files were preprocessed using R package minfi (vision 1.28.4) [34]. The sites which failed detection P = 0.01 were rewrote by nearest neighbor average to ensure enough number of sites.

**Comparation of the ability of discrimination**

For random forest prediction, we use R package randomForest (vision 4.6.14) and Number of trees are 5000[35]. 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 400[36]. The R package pROC (vision 1.14.0) was used to do ROC analysis to compare the abilities between hyper and hypo- sites by AUC[37].

**t-SNE analysis, PCA analysis and Gene Enrichment analysis**

tSNE analysis was performed by R package tsne (vision 0.1-3)[38]. PCA was performed by R function princomp and visualized by first two principal components. KEGG and GO enrichment were online analyzed by DAVID 6.8 (<https://david.ncifcrf.gov>)[39, 40], IPA also used for enrichment analysis for more elaborate result[41].

## 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 (Illumina 450K microarray) have been deposited to xxx. Other data involved in this study included GSE68060, GSE68838, GSE77954, GSE77965, GSE81211, GSE101764, GSE107352, GSE75546 and E-MTAB-6450. All the script involved in the study have been deposited to Github??

## **Author Contribution**

JF performed analyses, developed analysis methods and power calculations, interpreted results, and drafted the manuscript. JL xxxxx? SG reviewed and edited the manuscript. CT and YZ conducted array experiments. ZW collected and prepared tissue samples and collected results of clinical assays. SD enrolled patients and collected all the clinical information. DZ and JF 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) to D.Z., Major State Basic Research Development Program (2014CB542006), and the Key Research Program of the Chinese Academy of Sciences (KJZD-EW-L14) to C.Z.

## **Reference**

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA Cancer J Clin, 2018. **68**(1): p. 7-30.

2. Chen, W., et al., *Cancer statistics in China, 2015.* CA Cancer J Clin, 2016. **66**(2): p. 115-32.

3. Guo, S., et al., *Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.* Nat Genet, 2017. **49**(4): p. 635-642.

4. Wang, X., et al., *Hypermethylation reduces expression of tumor-suppressor PLZF and regulates proliferation and apoptosis in non-small-cell lung cancers.* FASEB J, 2013. **27**(10): p. 4194-203.

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

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.* Clin Epigenetics, 2014. **6**(1): p. 30.

7. Haikun Zhang, P.D., 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, W.M. and J.M. Carethers, *Genomic and epigenetic instability in colorectal cancer pathogenesis.* Gastroenterology, 2008. **135**(4): p. 1079-1099.

9. Hidaka, H., et al., *Comprehensive methylation analysis of imprinting-associated differentially methylated regions in colorectal cancer.* Clinical epigenetics, 2018. **10**(1): p. 150-150.

10. Shi, Y.X., et al., *Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer.* BMC Genomics, 2017. **18**(1): p. 901.

11. Lindqvist, B.M., et al., *Whole genome DNA methylation signature of HER2-positive breast cancer.* Epigenetics, 2014. **9**(8): p. 1149-62.

12. Raggi, C. and P. Invernizzi, *Methylation and liver cancer.* Clin Res Hepatol Gastroenterol, 2013. **37**(6): p. 564-71.

13. Morris, M.R. and F. Latif, *The epigenetic landscape of renal cancer.* Nat Rev Nephrol, 2017. **13**(1): p. 47-60.

14. Herman, J.G., et al., *Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers.* Cancer Research, 1995. **55**(20): p. 4525.

15. Kane, M.F., 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, 1997. **57**(5): p. 808.

16. Yoshiura, K., et al., *Silencing of the E-cadherin invasion-suppressor gene by CpG methylation in human carcinomas.* Proceedings of the National Academy of Sciences, 1995. **92**(16): p. 7416.

17. Witold, K., et al., *Adenomas - Genetic factors in colorectal cancer prevention.* Rep Pract Oncol Radiother, 2018. **23**(2): p. 75-83.

18. Rex, D.K., et al., *American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected].* Am J Gastroenterol, 2009. **104**(3): p. 739-50.

19. Perez-Silva, J.G., M. Araujo-Voces, and V. Quesada, *nVenn: generalized, quasi-proportional Venn and Euler diagrams.* Bioinformatics, 2018. **34**(13): p. 2322-2324.

20. Clemmensen, C., et al., *Gut-Brain Cross-Talk in Metabolic Control.* Cell, 2017. **168**(5): p. 758-774.

21. Straussman, R., et al., *Developmental programming of CpG island methylation profiles in the human genome.* Nat Struct Mol Biol, 2009. **16**(5): p. 564-71.

22. Swami, T. and H.C. Weber, *Updates on the biology of serotonin and tryptophan hydroxylase.* Curr Opin Endocrinol Diabetes Obes, 2018. **25**(1): p. 12-21.

23. Xiaolong, G., et al., *Intestinal Crosstalk between Microbiota and Serotonin and its Impact on Gut Motility.* Current Pharmaceutical Biotechnology, 2018. **19**(3): p. 190-195.

24. Berke, J.D., *What does dopamine mean?* Nat Neurosci, 2018. **21**(6): p. 787-793.

25. Deng, Y., et al., *Cloning and characterization of a novel human alcohol dehydrogenase gene (ADHFE1).* DNA Seq, 2002. **13**(5): p. 301-6.

26. Tae, C.H., et al., *Alcohol dehydrogenase, iron containing, 1 promoter hypermethylation associated with colorectal cancer differentiation.* BMC Cancer, 2013. **13**: p. 142.

27. Wang, D., et al., *IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data.* Bioinformatics, 2012. **28**(5): p. 729-30.

28. Qu, X., et al., *Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter.* Oncogene, 2016. **35**(50): p. 6403-6415.

29. consortium, B., *Quantitative comparison of DNA methylation assays for biomarker development and clinical applications.* Nat Biotechnol, 2016. **34**(7): p. 726-37.

30. Kang, K., et al., *A Genome-Wide Methylation Approach Identifies a New Hypermethylated Gene Panel in Ulcerative Colitis.* Int J Mol Sci, 2016. **17**(8).

31. Barrow, T.M., et al., *Smoking is associated with hypermethylation of the APC 1A promoter in colorectal cancer: the ColoCare Study.* Journal of Pathology, 2017. **243**(3): p. 366-375.

32. Damaso, E., et al., *Primary constitutional MLH1 epimutations: a focal epigenetic event.* Br J Cancer, 2018. **119**(8): p. 978-987.

33. Bormann, F., et al., *Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis.* Cell Rep, 2018. **23**(11): p. 3407-3418.

34. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-9.

35. Wiener, A.L.a.M., *Classification and Regression by randomForest.* R News, 2002. **2**: p. 18-22.

36. Ripley, W.N.V.a.B.D., *Modern Applied Statistics with S*. Fourth ed. 2002, New York: Springer.

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

38. Hinton, G.E., *Visualizing High-Dimensional Data Using t-SNE.* Journal of Machine Learning Research, 2008. **9**(2): p. 2579-2605.

39. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

40. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2009. **37**(1): p. 1-13.

41. Kramer, A., et al., *Causal analysis approaches in Ingenuity Pathway Analysis.* Bioinformatics, 2014. **30**(4): p. 523-30.

## **Figure and legends:**

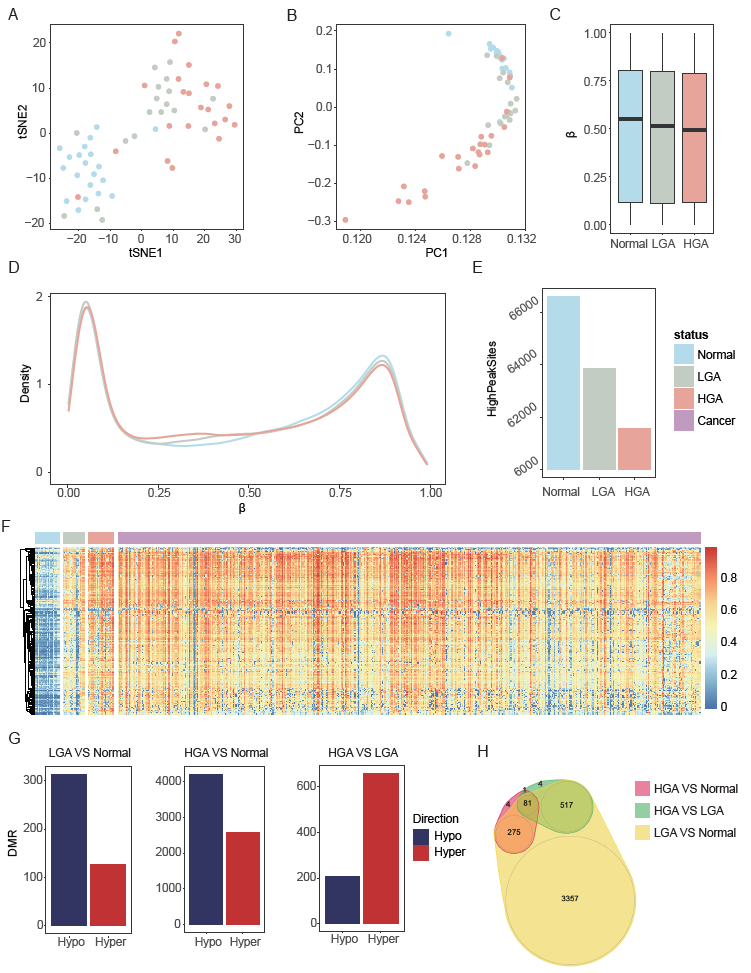


Figure 1. Genome-wide DNA methylation of low-grade adenoma (LGA), high-grade colorectal adenoma (HGA) and normal colorectal tissue. (A): tSNE analysis to show the data structure and sample relationship. (B): PCA analysis to show the data structure and sample relationship. (C): Average methylation level 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.

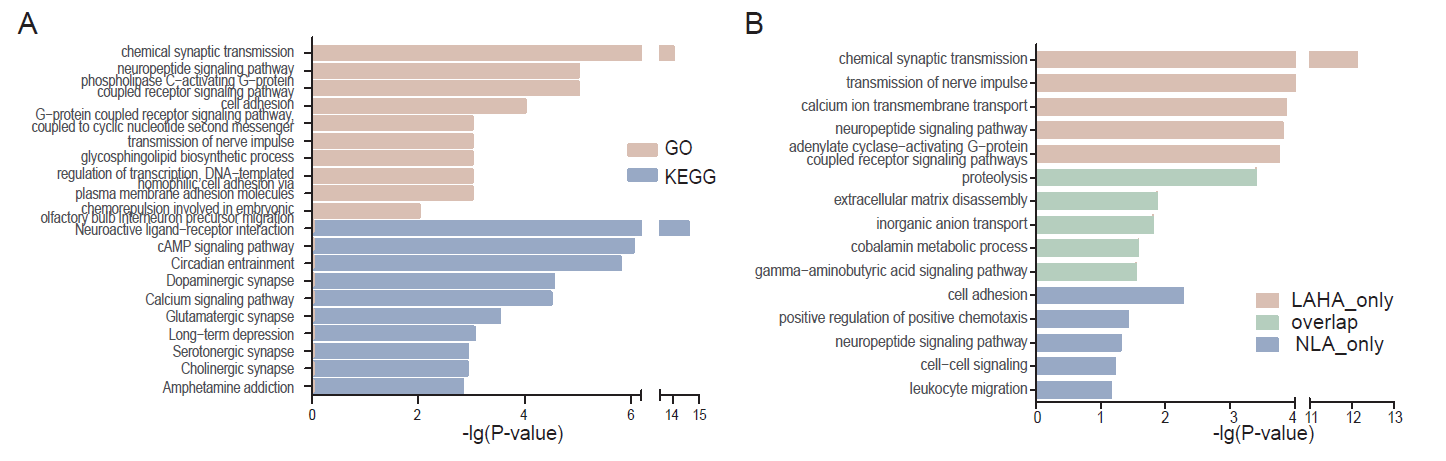


Figure 2. Enrichment analysis shown Nervous system was associated with adenoma development. (A) GO and KEGG analysis of the genes the LGA VS HGA DMRs located on. (B) GO analysis of the genes different DMRs located on, including the DMR only in LGA VS HGA, only in LGA VS Normal, and LGA VS HGA and LGA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

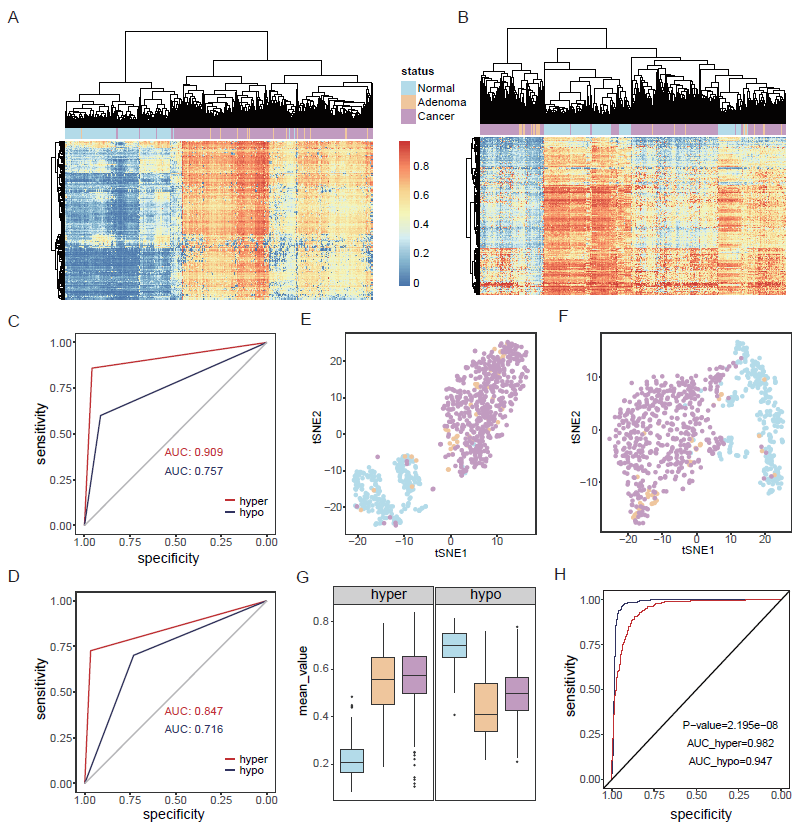


Figure 3. Hyper-methylated CpG sites showed better diagnostic performance than the hypo-methylated pattern. (A): Cluster analysis based on hyper-DMSs among normal, adenoma and cancer samples. (B): Cluster analysis based on hypo-DMSs among normal, adenoma and cancer samples. (C): Random forest prediction 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.

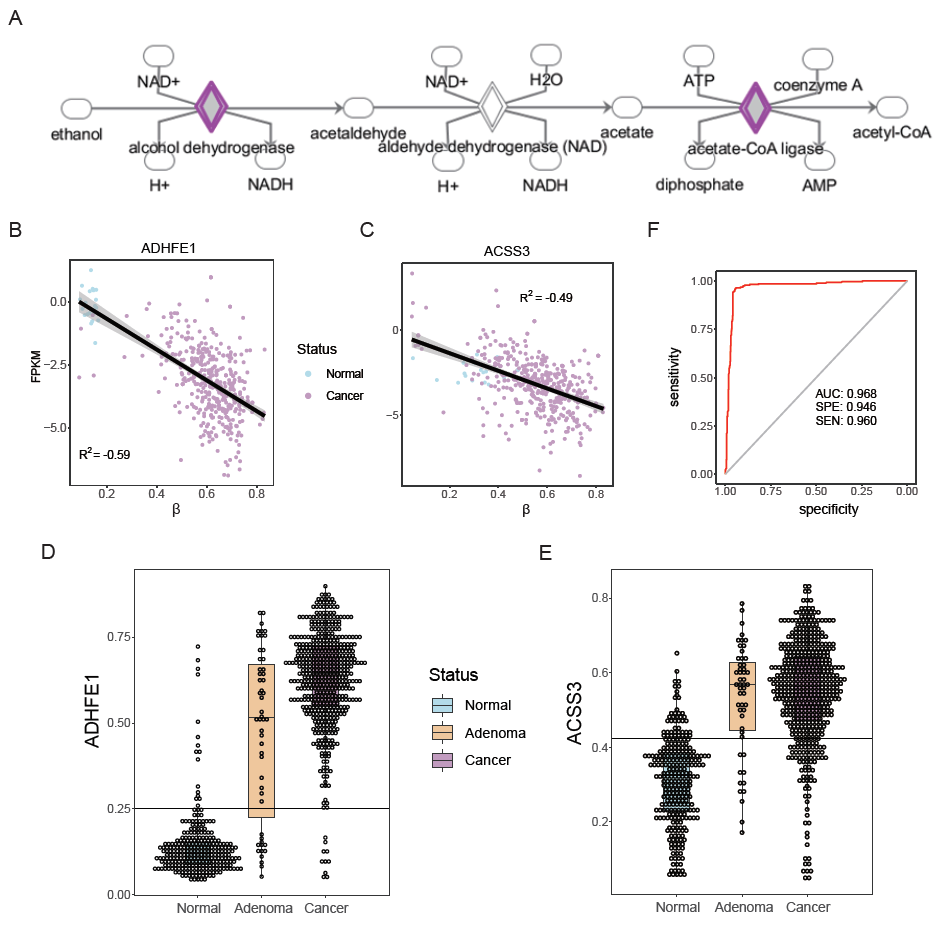


Figure 4. DNA methylation *ADHFE1* and *ACSS3* in Normal, LGA and HGA. (A): pathway of ethanol degradation II. (B): relationship between DNA methylation and gene expression of *ADHFE1*. (C): relationship between DNA methylation and gene expression of *ACSS3*. (D): DNA methylation of *ADHFE1* in normal adenoma and cancer samples. (E): DNA methylation of *ACSS3* in normal adenoma and cancer samples. (F): ROC of the prediction of *ADHFE1* for colorectal adenoma and caner.

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 |