**Genome-wide DNA methylation reveals methylation biomarkers for early diagnosis of colorectal cancer**

**Abstract**

1. **We have enough Figures, however, we need more tables to provide more details about the manuscript when we submit the paper.**
2. **Some other analysis should be conducted not only tSNE, but like random forest, SVM**
3. **I complete the revision for result section, and later I will revise introduction and discussion section later.**

**Introduction**

Colorectal cancer (CRC) is one of the major forms of cancer all over the world. Most colorectal cancers are due to old age and lifestyle factors, with only a small number of cases due to underlying genetic disorders. Colorectal cancer typically starts as a [benign tumor](https://en.wikipedia.org/wiki/Adenoma" \o "Adenoma), often in the form of an adenoma, which over time becomes [cancerous](https://en.wikipedia.org/wiki/Carcinoma" \o "Carcinoma). Colorectal adenoma is precursor lesion of colorectal cancer. In the colon, the evolution of normal epithelial cells to adenocarcinoma by and large follows a predictable progression of histological and concurrent epigenetic and genetic changes. In the classic colorectal cancer formation model, the vast majority of cancers arise from low grade adenoma, which then evolves into high grade adenoma before finally becoming a colorectal cancer. This process is driven by the accumulation of mutation and epigenetic alterations and take more than 10 years to occur.

Alterations of DNA methylation have been recognized as an important component of cancer development. Hypomethylation, in general, arises earlier and is linked to chromosomal instability and loss of imprinting. Global hypomethylation has also been implicated in the development and progression of cancer through different mechanisms. Generally, in progression to cancer, hundreds of genes are [silenced or activated](https://en.wikipedia.org/wiki/Regulation_of_transcription_in_cancer" \l "Transcription_silencing/activation_in_cancers" \o "Regulation of transcription in cancer). 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](https://en.wikipedia.org/wiki/CpG_site" \o "CpG site) in the [CpG islands](https://en.wikipedia.org/wiki/CpG_site" \l "CpG_island" \o "CpG site) that are present in the [promoters](https://en.wikipedia.org/wiki/Promoter_(genetics)" \o "Promoter (genetics)) of protein coding genes.

Over the past years, DNA methylation profiling has developed into an important approach for cancer research at molecular level. [Epigenetic](https://en.wikipedia.org/wiki/Epigenetics" \o "Epigenetics) alterations are much more frequent in colon cancer than genetic (mutational) alterations. Aberrant DNA methylation has been found in CRC and appears to also play an important role in driving CRC formation. The average CRC genome carries thousands of alterations in DNA methylation of CpG dinucleotides, often found on promoter regions of genes and inducing the transcriptional repression. CpG island methylator phenotype (CIMP) is a common feature of CRC, approximately 15%-20% CRC have it. Recently, a study found CIMP is a CRC specific feature, which is rare in colorectal adenoma. Another study found the normal tissue adjacent cancer show change of DNA methylation compared with the normal tissue from healthy people, which means change of DNA methylation level is a very early event. Adenoma always be considered as an intermediate state before CRC formation according classic adenoma-cancer module. Since adenoma removal significantly reduced the reduced the risk of colorectal cancer and most patients only need to resume surveillance colonoscopy after 5 to 10 years, limited studies describe changes in adenoma as early events during disease progression. Adenoma is not a simple status, which can be classified in two different stages, low grade adenoma and high grade adenoma, according to pathology. Just like CIMP is a colorectal cancer specific feature, adenoma is a precancerous lesion sharing similar feature as cancer as well as its own feature. So it is necessary to carry an epigenome-wide analysis for colorectal adenoma, in order bring a better understanding of CRC.

The mainstream view think DNA methylation of promoter region negatively influence gene expression. Moreover, hypermethylation and hypomethylation are two kind of different biochemical processes. The DNMTs can maintain DNA methylation or create de novo one, bringing about hypermethylation. The demethylation is considered by a cycle from 5mC, 5hmC, 5fC to unmethylation cytosine or passively rely on cell diversion, the processes take more complicated and indirect way back to unmethylation cytosine. The difference between hyper and hypo methylation illustrate we should consider them separately ever both of them regarded as DNA methylation variants.

Here, we conducted a study xx[describe what we have done with 1-2 sentences]xxxx and try to focus on colorectal adenoma, take hyper and hypo methylation as two biological processes, and try to illuminate the change on the benign stage.

**Results**

#### Whole genome hypomethylation occurs on precancerous benign lesion

We utilized the HM450 BeadChips array to profile DNA methylation on single-base level for low and high grade colorectal adenoma and adjacent normal tissue. We find the significant genome-wide DNA methylation difference between normal, low and high grade adenoma in the tSNE and PCA analysis (Figure xx). Compared with the adjacent normal tissue, low grade adenoma show whole genome hypomethylation (Pvalue = XXX, Supplementary Figure xx), and further hypomethylation occurs on high grade adenoma (Pvalue = XXX, Supplementary Figure xx). Both of two different stages of adenoma are benign lesion before cancerization. Bimodal distribution can characterize DNA methylation pattern, and hypermethylated peak can clearly reflect progressive hypomethylation (here histogram plot is better than density plot).

#### Different methylation regions (DMRs) in colorectal adenoma

We identified 440 DMRs in low grade adenoma compared with normal samples within gene associated regions including 126 (28.6%) hyper-methylated regions and 314(71.4%) hypo-methylated regions (Figure 2a). Methylation changes also found in high grade adenoma, a total 6,805 regions were differentially methylated compared with normal tissue including 2,592 (38.1%) hyper-methylated regions and 4,213(61.9%) hypo-methylated regions (Figure 2b). The hypo-methylation of the most DMRs in the low and high grade adenoma suggests global methylation change is an early event before colorectal cancer. To gain a better understanding of the dynamic methylation change of adenoma, we compared the methylation between high grade adenoma low grade adenoma and identified 868 DMRs in which 660 (76.0%) are hyper-methylated regions and 4,213 (24.0%) hypo-methylated regions (Figure 2c). [Here, we need a Figure to show the process and methylation in LAHA, NHA and NLA]. These result indicates DNA methylation started to be changed in the early stage of precancerous benign lesion including low and high grade adenoma and we found the significant distinct DMRs between the two stages indicating the different epigenetic process in these two stages (Figure 2d).

#### Nervous system is associated with adenoma development

We do enrichment analysis for 384 genes the DMRs located on, most terms are nervous system and signal transduction associated. Recent years, gut-brain cross-talk is focused by more and more studies, and dopaminergic synapse and serotonergic synapse are hit on KEGG enrichment result, which play a role in gut-brain axis model. 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 NLA and NHA without LAHA, and 571 significantly different methylated genes shown in LAHA and NHA without NLA. 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 development compared with occurrence. There still are some genes show methylation change in both disease occurrence and development stages. GO analysis shows the top term of these genes enriched for is cell adhesion, and positive regulation of positive chemotaxis and neuropeptide signaling pathway also be hit. [This paragraph is too vague and I cannot get the point. Hope to have certain update and then I can give some revision]

#### PRSS1 hypo-methylation is one of the earliest changes of adenoma

The genes significantly different methylated only from normal tissue to low grade adenoma were enriched for proteolysis (P=xx), extracellular matrix disassembly (P=xx), cobalamin metabolic process (P=xx) [Here, we need a Figure and a supplementary Table to list these genes]. PRSS1, ACAN, MMP12 were included in GO term of proteolysis and extracellular matrix disassembly while MMP26 was only included in proteolysis. Previous studies demonstrate these two biology process are significant associated with adenoma [reference?]. [proteolysis and extracellular matrix disassembly were associated with early change of adenoma?? Why mention these two terms here? We need evidence or result to support earliest change of adenoma here]. Meanwhile, in order to validate our hypothesis, we collected TCGA colon cancer dataset and found PRSS1 (P=xx), ACAN (P=xx), MMP12(P=xx) are significantly hypo-methylated in cancer samples compared with adjacent normal tissues? We also observed PRSS1 promoter region was significantly deceased in the process from normal to high adenoma (Figure 4a) and hypo-methylated in multiple cancers including COAD (P=xx), READ (P=xx) and PAAD (P=xx) (Figure 4b). Survival analysis shown low expression of PRSS1 was significantly associated with decreased overall survival (P=0.01) (**Figure 4c**). Overall, we demonstrated PRSS1 hypo-methylation could be one of most interesting early biomarkers for adenoma identification.

#### Hyper-methylated DMSs has better discrimination than the hypomethylated

In order to evaluation the distinguish ability of DNA methylation for normal tissue, adenoma and colon cancer tissue, we conducted several machine learning based prediction with DMRs identified in our dataset (Supplementary Table xx). We found both hyper-methylated loci and hypo-methylated loci could provide effective distinguish ability between normal and colon cancers which include our own x20? 30?x dataset and 832 public dataset collected from GEO database (Figure 3a and 3d). Meanwhile, we observed hyper-methylated sites can provided better distinguish between normal samples and the cancer samples in

unsupervised tSNE cluster analysis (**Figure 5b** and **Figure 5e**). In order to avoid the inconsistent result caused by unstable methylation based on single CpG site, we conducted same analysis based on mean beta value (mBV) of these sites. We found that the 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 5c). We observed similar results for hypo-methylation loci in which the average mBV of the normal tissue, the adenoma and the cancer are 0.698, 0.444 and 0.499 respectively (**Figure 5**). Finally, we applied xx model with hyper-mBV and hypo-mBV markers respectively and we found the area under the curve (AUC) of receiver operating characteristic (ROC) curve are 0.982 and 0.947, respectively. Permutation analysis based on bootstrap strategy shown the model based on hyper-methylated loci has better discrimination than the model of hypo-methylated loci (P<2.2x10-8, **Figure 5**).

#### 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 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 (P=xx), where two genes are hit, *ADHFE1* and *ACSS3*, which can facilitate translation form ethanol to ethanal and from acetic acid to acetyl-CoA respectively. 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=xx). We found the average methylation level of CpG loci located in CpG islands within ADHFE1, ACSS3 and CDO1 promoter region are significantly increased in cancer samples compared with normal samples (delta mBVs=0.2, 0.18 and 0.18 respectively, **Figure 6a**). Furthermore, we applied 24 CpG loci of ADHFE1, ACSS3 and CDO1 to distinguish the normal tissue and the disease tissues, and the tSNE analysis shown well discrimination performance (Figure 6b). Finally, we evaluated the discrimination power for these 3 genes separately. The result shown ADHFE1 has better discrimination power compared with ACSS3 and CDO1 (Figure 6c). [Here, tSNE is only non-supervised analysis, we need supervised prediction analysis and provided AUC curve for each biomarker and biomarker recombination, such as random forecast, SVM or logistic regression model]

**Methods**

**Sample collection and pathologist confirmation**

The normal tissue and adenoma tissue specimens were collected from the patients who underwent endoscopic treatment in the Department of Gastroenterology of Peking University Third hospital between March2015 and June 2016.Tissue specimens were embedded in paraffin, sectioned, stained with hematoxylin and eosin, and confirmed by pathologist by light microscopy.

**DNA isolation and bisulfite conversion**

**（等程程）**

**Methylation data processing**

In order to investigate the epigenomic landscape of colorectal adenoma, the HM450 BeadChips array was used to identify DNA methylation changes in low and high grade colorectal adenoma from adjacent normal tissue. DMRs defined as genomic region exhibiting an average change in methylation > 15%, namely the average change of beta value > 0.15, compared to adjacent normal tissue that achieved a false discovery rate < 0.05, were used for all further analysis (Method). The HM450 array categorize probes based on gene regions into 3 major gene feature groups: promoter (5’UTR, TSS200, TSS1500 and 1stEXON), intragenic regions (GENEBODY and 3’UTR) and intergenic regions.

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 locus 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

Methylation levels of different regions including TSS1500, TSS200, 5’UTR, 1stEXON, GENEBODY and 3’UTR, were identified by IMA, then different methylation region were defined by rank sum test following FDR adjust P value<0.05 and different methylation level>0.15. Different methylation sites were defined by rank sum test following FDR adjust P value<0.05 and different methylation level>0.2.

**Public data collection and processing**

GSE68060, GSE68838, GSE77954, GSE77965, GSE81211, GSE101764, GSE107352 and GSE75546 were collected from GEO, E-MTAB-6450 was collected from ArrayExpress. All of these datasets accessing raw data idat files, were preprocessed using R package minfi. The sites which failed detection P value were rewrote by nearest neighbor averaging.

**Discussion**

Whole genome DNA hypomethylation and promoter hypermethylation of cancer related gene are regard as the common pattern of cancer. The most DMRs of the low grade adenoma are hypomethylated compared with the normal tissue, and we also obverse whole genome hypomethylation of low grade adenoma which is consistent with previous reports of colorectal cancer. Taking IPA results into consider, the main terms are related to immuno-inflammatory response, like (a)granulocyte adhesion and diapedesis, inhibition of matrix metalloproteases and so on. We speculate whole genome hypomethylation may be correlated with immunity, which is a very early manifestation on colorectal adenoma. While the most DMRs of the high grade adenoma are hypermethylated, the main terms of IPA are related to nervous system. It is hard to explain the results, we guess it may related to gut-brain axis, still needs further study to figure it out.

We classified DMRs into hyper and hypo methylation group and found hypermethylated DMRs are enriched on promoter region, consistent with the previous report. Whereas, hypomethylated DMRs are enriched on gene body and 3’UTR regions, just when compared with the hyper. The 3’UTR significantly enriched by hypomethylated DMRs, which is a new discovery. But gene body region doesn’t show enrichment when compared with the region distribution of whole array.

Focusing on hypo-methylation part, 3 MMPs gene in the term inhibition of matrix metalloproteases, one of which, MMP12, show increasing expression on colonic and rectal cancer tissue compared to the normal tissue. Subsequently, we check every site on MMP12 promoter region, there are two sites harbor relatively greater DNA methylation difference, the different methylation levels of cg20487452 and cg23979520 are -0.160 and -0.171, with P value 8.32e-06 and 1.33e-06. Similarly, we get 2 potential sites of cg23684449 and cg04839706 whose different methylation levels are -0.172 and -0.190, with P value 4.64e-06 and 3.87e-05 on GPT2 promoter, the gene is the only gene enriched in alanine degradation and biosynthesis.

We also use the 4 sites of MMP12 and GPT2 to distinguish the normal tissue and the disease tissues, and the tSNE plot also show the ability of discrimination. Beyond that, we use 832 samples of public data as a validation set helps to confirm the ability of discrimination to some degree. It’s obvious that the 24 sites from hypermethylated DMRs perform better than the 4 sites from hypomethylated DMRs

Meanwhile, PRSS1 was also found in cobalamin metabolic process. Cobalamin, also called Vitamin B12, is a [cofactor](https://en.wikipedia.org/wiki/Cofactor_(biochemistry)" \o "Cofactor (biochemistry)) in DNA synthesis, and in both fatty acid and amino acid metabolism, specially which is the only one vitamin absorbed by help of intestinal secretions (gastric internal factor). PRSS1 has been reported downregulated on pancreatic adenocarcinoma (PAAD), and is also related with hereditary pancreatitis. The TSS1500 and TSS200 of PRSS1 are significantly hypomethylated on adenoma compared with the adjacent normal tissue. For further study, we queried TCGA data. The gene significantly downregulated on colonic and rectal cancer tissue. Besides, it has significant survival difference on expression level, and low expression is associated bad survival.

Totally, we identified 440 DMRs between the normal tissue and the low grade adenoma, except TSS1500 of CDX2, 439 DMRs maintain or strengthen its methylation difference at the high grade adenoma. The DMRs identified on the low grade adenoma show highly robust on further status, which is consistent with the classic CRC model, to some degree reminds us early event is vital. Similar as the big bang model, at epigenetic level, our result also underline the key role of early events. CDX2 a member of the [caudal](https://en.wikipedia.org/wiki/Caudal_(protein)" \o "Caudal (protein)) related [homeobox](https://en.wikipedia.org/wiki/Homeobox" \o "Homeobox) [transcription factor](https://en.wikipedia.org/wiki/Transcription_factor" \o "Transcription factor) family that is expressed in the nuclei of intestinal [epithelial cells](https://en.wikipedia.org/wiki/Epithelial_cell" \o "Epithelial cell). The TSS1500 of CDX2 show reverse methylation change on the high grade adenoma, it is hard to verdict this is a biological phenomenon or data bias, more sample needed to confirm it.

We identified DMSs between the normal tissue and the low grade adenoma, divide the DMSs into hyper and hypo group. The public data from GEO an ArrayExpress show hypermethylated DMSs has better discrimination for normal tissue and disease tissue compared with hypomethylated DMSs. The mBV can estimate the average level of sites in the same group, we do plot the ROC curves for the hyper and the hypo mBV, and the result also show the bigger AUC of hyper mBV.

In order to figure out the potential function difference between the hypermethylation and hypomethylation, we separate DMRs between the normal tissue and the low grade adenoma into hyper and hypo DMRs and followed by the IPA enrichment analysis. Hypomethylated DMRs enriched for inhibition of matrix metalloproteases, alanine degradation and biosynthesis, bladder cancer signaling, granulocyte adhesion and diapesis and so on, the most of them are immune-inflammatory response associated, the most of them are routine biological processes. Hypermethylated DMRs enriched for ethanol degradation and taurine biosynthesis and so on, the most of them are response to the external ingesta like ethanol, and taurine mainly intaking from external system directly. Notably, even there are some other terms, which share same or similar gene with top 2 terms.

Finally, we try to estimate potential of the DMRs as biomarker, taking IPA results into consideration, to distinguish the normal tissue from disease tissue. we check the expression level of gene enriched on the top terms of IPA results, ACSS3, ADHFE1, and CDO1 show down expression at cancer tissue. MMP12 and GPT2 show up expression at cancer tissue. Then, we pick 4 sites from MMP12 and GPT and 24 sites from ACSS3, ADHFE1, and CDO1 respectively, to distinguish the normal tissue and cancer tissue. the 24 hypermethylated sites harbor better discrimination then the 4 hypomethylated sites. The result further illustrate hypermethylation may play more important role during the disease development. And the hypermethylation may be influenced by the external system.

ACSS3, ADHFE1 and CDO1 make contributions to high separability between the normal tissue and disease tissue. But the 3 genes contribute as a whole or the certain one playing the key role is not clear. The tSNE plot show ADHFE1 can make similar separability as the 3 genes together. 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. Our study first report ADHFE1 DNA methylation difference at colorectal adenoma, even the low grade.

The result goes the DNA methylation change of the gene is a very early event, and give the gene potential as a DNA methylation biomarker of colorectal adenoma and cancer.

**Disclosure of Conflicts of Interest**

The authors declare no conflict of interest.

**Abbreviations**

**Figure Legends:**

Figure 1,

Figure 2,

Figure 3,

Figure 4,

Figure 5,