RESEARCH

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

Jian Fan1,6†, Jun Li2†, Shicheng Guo3,4†, Chengcheng Tao1, Haikun Zhang1,6, Wenjing Wang2, Ying Zhang1, Dake Zhang1,5*, Shigang Ding2* and Changqing Zeng1*

Abstract

Background: Abnormal DNA methylation is a hallmark of human cancers and may be a promising biomarker for early diagnosis of human cancers. However, the majority of DNA methylation biomarkers that have been identified are based on the hypothesis that early differential methylation regions (DMRs) are maintained throughout carcinogenesis and could be detected at all stages of cancer.

Methods: In this study, we identified potential early biomarkers of colorectal cancer (CRC) development by genome-wide DNA methylation assay (Illumina infinium450, 450 K) of normal (N = 20) and pre-colorectal cancer samples including 18 low-grade adenoma (LGA) and 22 high-grade adenoma (HGA), integrated with GEO and ArrayExpress datasets (N = 833).

Results: We identified 209 and 8692 CpG sites that were significantly hyper-methylated in LGA and HGA, respectively. Pathway analysis identified nervous system-related methylation changes that are significantly associated with early adenoma development. Integration analysis revealed that DNA methylation in the promoter region of ADHFE1 has the most potential for being an early diagnostic biomarker for colorectal adenoma and cancer (sensitivity = 0.96, specificity = 0.95, area under the curve = 0.97).

Conclusions: Overall, we demonstrated that DNA methylation have been shown significant changes in the stage of LGA and HGA in the development of colon cancer. Genome-wide DNA methylation to LGA and HGA provided an important proxy to identify promising early diagnosis biomarkers for colorectal cancer.

Keywords: DNA methylation, Low-grade adenoma, High-grade adenoma, Colorectal cancer, Biomarker

© The Author(s). 2020 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.
Background

Colorectal cancer (CRC) is the third leading cause of cancer-related deaths worldwide [1, 2]. Current evidence indicates that genetic mutations and epigenetic alterations progressively accumulate in the tumor genome during carcinogenesis, and these alterations may serve as primary biomarkers for early detection and treatment of cancer [3]. Abnormal alterations in methylation status specifically hyper-methylation or hypo-methylation have also been associated with abnormal tissue differentiation. Altered methylation has been observed in the promoter regions of tumor suppressor genes and miRNA have been observed in almost all cancer types [4, 5]. Over the past decades, detection of altered DNA methylation has been widely studied to develop cancer biomarkers [6] and the majority that have been identified are based on the hypothesis that early differential methylation regions (DMRs) are maintained throughout carcinogenesis and could be detected at all stages of cancer. For example, altered methylation patterns have been detected with hepatic disease progression in the context of hepatitis, cirrhosis, and hepatocellular carcinoma (HCC) [7]. Moreover, recent evidence demonstrated that cell-free DNA (cfDNA) methylation can be used for early cancer diagnosis and tissue-of-origin mapping for metastatic cancer [4].

Abnormal alterations of DNA methylation have been recognized as an important event in cancer development [8]. Global hypo-methylation arises early in carcinogenesis and has been linked to chromosomal instability and loss of imprinting [9, 10]. Generally, during cancer development, hundreds of genes are silenced or activated [11–13]. Although silencing of some genes in cancers occurs by mutation, a large proportion of carcinogenic gene silencing is a result of altered DNA methylation [14]. DNA methylation-based silencing in cancer typically occurs at multiple CpG sites in the CpG islands present in the promoters of protein-coding genes [15]. Although extensive epigenetic alterations have been defined over the past years, CRC is still not well understood at the molecular level [3]. Against a background of whole-genome hypo-methylation, gene-specific promoter hyper-methylation has been found to promote CRC by downregulating the expression of key tumor suppressor genes such as CDKN2A, MLH1, and CDH1 [16–18]. CRC is a heterogeneous disease that typically originates from a pre-cancerous lesion, often in the form of an adenoma, eventually progressing to a malignant cancer within a temporal window that may exceed 10 years [19]. Because CRC exceeds many other cancers in both incidence and mortality, capacity to detect and monitor molecular changes during the colorectal adenoma (AD) stage provides an excellent opportunity to prevent cancer progression and improve survival outcomes [20]. While a large number of studies have focused on CRC, a subset of them has focused specifically on the adenoma as an intermediate stage which required more specific molecular definition. For instance, a ten-gene methylation signature in adenoma was found to change with pathologic progress [21]. Notably, colorectal adenoma has two pathologic stages: low-grade adenoma (LGA) and high-grade adenoma (HGA) [22]. Our research compared and defined differences of genome-wide profiling of DNA methylation, especially changes across these two pre-cancerous stages that had not been reported [23]. We hypothesized that these alterations in LGA methylation represented candidates as potential early diagnostic biomarkers. We further posit that comprehensive understanding of the genome-wide DNA methylation profile for 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 conducted a series of genome-wide DNA methylation array of 18 LGA and 22 HGA and compared the frequency, location, and pattern of methylation status of 20 normal tissue samples. Dynamic DNA methylation changes were identified for LGA and HGA, and we found that methylation changes that appeared in LGA were increased or maintained in HGA and cancer. Enrichment analyses to DMRs were performed to further investigate the potential influence of DNA methylation on functional difference in adenoma initiation and development. Moreover, we separated different methylation sites (DMSs) between LGA and normal into hyper-DMS and hypo-DMS and evaluated their respective performance for CA and CRC prediction. To validate our findings, we compared them to genome-wide DNA methylation profiles of 833 samples from public database. Finally, we describe the identification and analysis of one functional methylation signature at the promoter region of ADHFE1 as a potential biomarker for early CRC development.

Results

Landscape of DNA methylation of pre-cancerous lesions

We profiled DNA methylation at the single-base level for 18 LGA, 22 HGA, and 20 normal tissues. We found significant genome-wide DNA methylation differences among normal-, low-, and high-grade adenoma (Fig. 1a, b). Compared to normal tissue, LGA had genome-wide hypo-methylation ($P = 5.2 \times 10^{-5}$, rank sum test) which was even lower in HGA ($P = 3.7 \times 10^{-6}$, compared with normal, rank sum test, Fig. 1c). Methylation levels of all target sites in the array demonstrated the known bi-modal distribution in normal, LGA, and HGA (Fig. 1d), and the amount of fully methylated sites of lesions decreased with increasing degree of malignancy (right
Fig. 1 (See legend on next page.)
peak, Fig. 1d, e). Almost all DMSs in LGA compared to normal tissues kept at least an equivalent methylation level if not higher than in HGA and cancer (Additional file 1: Fig. S1). The 209 significantly hyper-methylated sites in LGA were further hyper-methylated in 22 HGA and 504 cancer samples collected from public databases (Fig. 1f and Additional file 1: Fig. S2, Table S1), and hypo-DMSs had a diametric tendency (Additional file 1: Fig. S3) suggesting that DNA demethylation may occur very early in pre-cancerous lesions. Over 60% of DMRs that were observed in both LGA (71.4%, 314/440) and HGA (61.9%, 4,213/6,805) were hypo-methylated compared to normal tissues (Fig. 1g, Additional file 1: Table S2 and S3). However, with LGA as the reference, most DMRs observed in HGA were hyper-methylated (76.0%, 660/868) (Fig. 1g, Additional file 1: Table S4). In addition, there were limited overlaps between genes with DMRs in LGA compared to normal tissues and those compared to HGA, suggesting different epigenetic process (Fig. 1h) [24].

Nervous system processes were associated with adenoma development

Enrichment analysis of 603 DMRs which were located between HGA and LGA, and most highly enriched functional terms, included the nervous system and those associated with signal transduction (Fig. 2a), specifically dopaminergic synapse and serotonergic synapse pathways, which play a role in the gut–brain axis model of signaling cross-talk between organ systems [25]. These results correspond to gene methylation findings in Fig. 1g where HGA vs normal includes almost all genes that are listed in LGA vs normal and HGA vs LGA DMRs. To figure out the potential function changes from LGA to HGA, Gene Ontology (GO) enrichment was performed for 275 genes that were significantly different in methylation status between LGA vs normal and HGA vs normal without considering the differences in methylation status between HGA vs LGA. Five hundred seventy-one significantly different methylated genes were highlighted in HGA vs LGA and HGA vs normal without LGA vs normal (Fig. 2b). For the 275 genes with significantly different methylation patterns in only the LGA vs normal and HGA vs normal comparisons, GO analysis selected the top enriched terms of proteolysis as well as extracellular matrix disassembly, inorganic anion transport, and cobalamin metabolic processes. Cell adhesion, positive regulation of positive chemotaxis, and neuropeptide signaling pathway were term hits on the overlapping part between LGA vs normal and HGA vs LGA. Genes that were significantly different in methylation status between LGA and HGA were enriched for chemical synaptic transmission, transmission of nerve impulse, calcium ion transmembrane transport, and similar neural processing terms. Like the DMR enrichment analysis, terms related to the nervous system were selected yet exhibited different term patterns between HGA vs LGA compared to LGA vs normal.

Hyper-methylated CpG sites exhibited better discrimination between normal, pre-cancerous, and cancerous tissues than the hypo-methylated pattern for CRC

To distinguish the discriminatory ability of DNA methylation patterns for normal tissue, CA, and CRC, we collected 833 genome-wide DNA methylation datasets from GEO and ArrayExpress, public datasets which included 278 normal tissue samples, 51 adenoma samples, and 504 cancer samples. We separated DMSs of LGA vs normal into two groups including hyper-DMSs (209 sites) and hypo-DMSs (441 sites). We found both hyper-DMSs and hypo-DMSs could effectively distinguish methylation pattern differences between disease (adenoma and cancer) and normal samples (Fig. 3a, b). Meanwhile, we also conducted two machine learning-based predictions with the DMSs identified in our dataset and observed that hyper-methylated sites can better distinguish between normal samples and disease samples via random forest and neural network methods (Table 1). For hyper-methylated sites, the area under the curves (AUCs) of receiver operating characteristic (ROC) curves were 0.91 and 0.85, respectively. For hypo-methylated sites, AUCs of ROC curves were lower at 0.72 and 0.76, respectively (Fig. 3c, d). Unsupervised t-SNE cluster analysis produced the same result (Fig. 3e, f). To avoid inconsistent results caused by unstable methylation based on single CpG sites, we compared the mean beta value (mBV) of these sites. We found that hyper-methylated mBVs were significantly different between normal tissue and CRC ($P < 2.2 \times 10^{-16}$); however, there was no significant difference between the adenoma and cancer ($P = 0.29$, Fig. 3g) in which the average mBV of the normal tissue, adenoma, and 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, adenoma, and cancer were 0.70, 0.44, and 0.50, respectively (Fig. 3g). Finally, we found the AUCs of ROC curves with hyper-mBV and hypo-mBV were 0.98 and 0.95, respectively. Permutation analysis based on a bootstrap strategy indicated that the model based on hyper-methylated sites had better discriminatory power than the model of hypo-methylated sites (P < 2.2 × 10⁻⁸, Fig. 3h).

The promoter of ADHFE1 may be a potential biomarker for colorectal adenoma and cancer

Next, we grouped the DMRs of normal tissue and LGA into hyper- and hypo-DMRs and performed enrichment analysis by Ingenuity Pathway Analysis (IPA). The top enriched functional term for hyper-DMRs was ethanol degradation II (P = 5.4 × 10⁻³) which was mostly contributed to methylation sites on two genes, ADHFE1 and ACS3, which can facilitate the conversion from ethanol to acetaldehyde and from acetic acid to acetyl-CoA, respectively (Fig. 4a). The expression of both genes were downregulated in colonic and rectal cancer tissue compared with normal tissue (P < 0.01), a result consistent with the DNA methylation changes between LGA and HGA (R² = -0.49 and -0.59, Fig. 4b, c). We found that the average methylation level of CpG sites located in CpG islands within the promoter regions of ADHFE1 and ACS3 were significantly increased in cancer samples compared to normal samples (ΔmBVs = 0.2 and
0.18, respectively. We further analyzed the promoter region within the CpG island of the two genes to distinguish between normal and disease tissues. When setting the cutoff at 0.25 for the \textit{ADHFE1} promoter, the minimal error rate was only 4.68% (39/833, Fig. 4d); the heatmap of sites within the region reflected the same result (Fig. 4e). ROC curve analysis of mBV of the \textit{ADHFE1} promoter for all 833 samples produced an AUC of 0.97 with specificity and sensitivity at 0.95 and 0.96 (Fig. 4f). For cancer samples, an AUC as high as 0.98 was determined (Additional file 1: Fig. S4). For ACSS3, the minimal error rate of its promoter was
16.68% (139/833) with a cutoff set at 0.42 (Fig. 4g) which performed inferiorly to ADHFE1 in terms of discrimination power. Meanwhile, we also compared ADHFE1 with SEPT9, an FDA-approved methylation-based biomarker for CRC screening. The correlation of the two genes was 0.77, and we determined that ADHFE1 had a better prediction power than SEPT9 (Fig. 5a and Additional file 1: Fig. S5) [27]. Furthermore, we observed ADHFE1 to have a much better separation boundary compared to SEPT9 (Fig. 5b). In view of most detected cfDNA being actually the fragments from white blood cells, we checked DNA methylation status of ADHFE1 promoter in 656 whole blood cases from public data. As expected, all sites in the promoter were consistently at low methylation level (Additional file 1: Fig. S6).

**Discussion**

Whole-genome DNA hypo-methylation and hyper-methylation analysis of the promoter regions of cancer-related genes is regarded as a common method of characterizing diverse cancers [28]. In our study, we found that whole-genome DNA hypo-methylation may start at the LGA stage and lead to further hypo-methylation at HGA and CRC (Fig. 1c). As many previous studies have reported, a bimodal distribution can characterize DNA methylation pattern, and we noted that the hyper-methylated peak can clearly reflect progressive hypo-methylation (Fig. 1d, e) [29]. We identified 440 and 6805 DMRs in low- and hyper-grade adenoma, respectively, and of these DMRs, 314 (71.4%) in LGA and 4213 (61.9%) in HGA were hypo-methylated compared to normal tissues. On the contrary, most DMR (660/868, 76.0%) differences between HGA and LGA were hyper-methylated. Aside from a little overlap between HGA genes, significantly distinct DMRs were located between LGA vs normal and HGA vs LGA which indicates that LGA vs normal and HGA vs LGA are possibly not the same process with a degree difference but two different epigenetic processes. These genome-wide demethylation patterns may indicate that though hypo-methylation dominates the carcinogenesis of CRC, hyper-methylation sites may contribute more to the distinct malignancy of these lesions.

To find functional differences between differing methylation patterns in normal, pre-cancerous, and cancerous tissues, enrichment analysis was applied to 603 genes with DMRs between HGA and LGA which determined that most enriched terms were related to nervous system and signal transduction (Fig. 2a). The term gut–brain-axis describes an integrative physiology concept that incorporates all, including afferent and efferent neural, endocrine, nutrient, and immunological signals, cross-talk between the central nervous system, and the gastrointestinal system that may be dysregulated during carcinogenesis [25]. Our Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis further highlighted the significance of dopaminergic synapse and serotonergic synapse to CRC development. Serotonin (5-hydroxytryptamine; 5-HT) is popularized as a contributor to feelings of well-being and happiness though its actual biological function is complex and multifaceted with roles in modulating cognition, reward, learning, memory, and numerous physiological processes [30]. Brain 5-HT gets much more respect, and certainly more press and research, than the vastly larger store of 5-HT in the gut though both are important for physiological functions [31]. Dopamine (3,4-dihydroxyphenethylamine; DA) is an organic chemical of the catecholamine and phenethylamine families that functions as both a hormone and a neurotransmitter and plays several important roles in the brain and body [32]. In the brain, dopamine functions as a neurotransmitter to send signals to other nerve cells [32]. Outside the central nervous system, dopamine functions primarily as a local paracrine messenger to reduce gastrointestinal motility and protect the intestinal mucosa [32]. The interaction of tumor and the nervous system has also been found in gastric cancer and liver cancer [33, 34]. Our study suggests that the gut–brain axis and related molecules may be important contributors to the
Fig. 4 (See legend on next page.)
development and progression of CRC even at the adenoma stage.

DNA methylation has always been considered as a potential biomarker for many diseases due to its tissue specificity and stability [35]. Here, we analyzed DNA methylation patterns as a mechanism to distinguish disease samples (including adenoma and cancer) from normal samples during CRC development. We identified 209 hyper-methylated sites and 441 hypo-methylated sites between LGA and normal samples and noted that both hyper- and hypo-methylated sites could effectively distinguish between normal and CRC tissues. Further validation with random forest and neural network analyses confirmed our observations. Specifically, AUCs of ROC curves for our prediction model using hyper-methylated sites were larger than those using hypomethylated sites, despite the observation that hyper-methylation at several key sites may trigger widespread hypomethylation throughout the genome during cancer development.

Colorectal adenoma is considered the middle stage between normal status and cancer; therefore, our study focused on identifying and comparing the differences in DNA methylation patterns among normal, pre-cancerous, and cancerous colorectal tissues. IPA enrichment analysis of hyper-DMRs identified in very early stage cancers selected Ethanol degradation II as the top term for functional impact, in which ADHFE1 and ACSS3 were hit. Intense early changes in DNA methylation patterns at the promoter region of these genes support their potential use as adenoma biomarker. It is known that ADHFE1 encodes for hydroxyacid-oxoacid transhydrogenase which is responsible for the oxidation of 4-hydroxybutyrate in mammalian tissues [36]. Some studies have also reported that the gene is associated with cell proliferation and differentiation [36–38]. In CRC tissue, ADHFE1 is hyper-methylated in the promoter region corresponding to downregulation of expression that may facilitate tumor growth [38]. Our results suggest that the DNA methylation of the ADHFE1 promoter is a potential biomarker for distinguishing colorectal adenoma and cancer from normal tissue.

As the only FDA-approved liquid biopsy marker for DNA methylation, SEPT9 has been applied for colon cancers screening [39]. Actually the detection signal of SEPT9 has been shown to be more distinguishable in tissues than at cfDNA samples [40]. The better performance of ADHFE1 than SEPT9 at tissue level made it a promising liquid biopsy biomarker for CRC. Further efforts with a larger, more diverse sample population are needed to validate the predictive efficacy of this biomarker at cfDNA.

In addition, a recent study found a promising biomarker cg10673833 which distinguished tumor patients from healthy people by cfDNA [41]. However, the methylation level of this marker showed only a slight upward trend from normal tissues to adenoma and cancer, in our samples as well as in public data.
In view of the very low methylation of cg10673833 in the blood, most likely its detection of cancer was mainly due to largely increased metabolism of the tumor tissue that caused increased shedding of ctDNA. Comparing with cg10673833, the better discrimination of normal to adenoma and cancer by ADHFE1 raises a great potential for this candidate as a methylation marker to indicate pathological changes.

Besides ADHFE1, we obtained a group of 209 hypermethylated DMSs in our LGA samples. For their potential being candidates of methylation markers, we examined these sites in 656 cases of whole blood from GEO. As shown in the heatmap of Additional file 1: Fig. S7, 207 out of 209 sites showed their low methylation level as < 0.3 in average, implying the potential of these sites deserving further investigation for early diagnosis.

Conclusions
Adenoma samples are perfect proxy for colorectal carcinoma early biomarker identification. Our study focused on adenoma, in order to get the earliest clue to detect colorectal disease. DNA methylation is a promising biomarker for cancer diagnosis and surveillance for its tissue specificity and robustness. We established the DNA methylation landscape of LGA and HGA and noted the hyper-methylated peak has a regular decrease accompanied with disease procession. Furthermore, we found the development of adenoma is associated with functions of nervous system, while the initiation of the adenoma is more associated with cell biological functions. Another relatively independent work was based on the precious finding in LGA, in which we found ADHFE1 is a potential early diagnosis biomarker of colorectal carcinoma and adenoma. Eight hundred thirty-three samples from the public database strongly support the gene is a promising biomarker.

Methods
Sample collection and pathological confirmation
In the Department of Gastroenterology of Peking University Third hospital from March 2015 to June 2016, we collected 18 LGA and 22 HGA specimens from patients who underwent endoscopic treatment for CA removal and obtained adjacent normal tissue specimens from 20 patients with adenoma during the treatment. Tissue specimens were embedded in paraffin, sectioned and stained with hematoxylin and eosin, and confirmed by pathologist by light microscopy. All the patients were treatment naive before their surgeries. Clinical information of patients, and sample position in corresponding microarray are provided in Additional file 1: Table S5 and S6.

DNA isolation and bisulfite conversion
DNA was isolated using QIAmp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer’s protocol. Bisulfite conversion was performed using the EZ DNA Methylation-Gold Kit according to the instruction manual (Zymo Research, Irvine, CA, USA).

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 and covers 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 raw data from the array was processed using the GenomeStudio Methylation (version 1.8, Illumina) module which calculated methylation levels. The GenomeStudio is the software for array data processing of Illumina, which integrates data normalization, background adjustment, and methylation calculation. Normalization was performed by comparing control probes when the option was set as controls, and background adjustment was performed automatically by the software selecting Subtract Background. The distribution of beta values before and after normalization across all was analyzed (Additional file 1: Fig. S8), and multi-dimensional scaling (MDS) according to 10,000 most variable positions showed the homogeneity of samples and their clustering according to pathological groups. Beta MDS were also analyzed according to 1000 and 20,000 most variable positions for all samples before and after normalization (Additional file 1: Fig. S9). The methylation status for each CpG site was calculated as the ratio of fluorescent signals ($\beta = \frac{\text{Max}(M,0)}{\text{Max}(M,0) + \text{Max}(U,0) + 100}$), ranging from 0 to 1 using the average probe intensity for the methylated (M) and unmethylated (U) alleles. $\beta = 1$ indicates complete methylation; $\beta = 0$ represents no methylation. Probes located on sex chromosomes or failed detection $P$ value testing of at least one sample or SNP (single-nucleotide polymorphism) were removed from analysis using R package IMA (vision 3.1.2) [42]. DMRs were defined as rank sum test following false discovery rate (FDR) adjusted $P$ value $< 0.05$ and $|\Delta \beta| > 0.15$, and DMSs were defined as rank sum test following FDR adjusted $P$ value $< 0.05$ and $|\Delta \beta| > 0.20$. Promoter regions were defined as 5’UTR, TSS200, TSS1500, and first exons.

Public datasets and processing
To ensure consistency of data processing, we only compared our samples with publically accessible samples with raw idat files. GSE68060, GSE68838, GSE77954,
GSE77965, GSE81211, GSE101764, GSE107352, and GSE75546 were collected from GEO while E-MTAB-6450 was collected from ArrayExpress [43–48] (Additional file 1: Table S6). Some cell line samples and metastatic cancer samples were removed upon further study. In total, we collected 278 normal samples, 51 adenoma samples, and 504 cancer samples. All datasets using raw idat files were preprocessed using R package minfi (version 1.28.4) [49]. The sites which failed detection at $P = 0.01$ were rewritten to the nearest neighbor average to ensure an adequate number of sites for analysis. Six hundred fifty-six cases of whole blood data were collected from GEO (accession number GSE40279).

Comparison of the ability of discrimination between normal, LGA, HGA, and CRC tissue
For random forest prediction, we used R package randomForest (version 4.6.14) with the number of trees set at 5000 [50]. For neural network prediction, we used R package nnet (version 7.3.12) with number of units in the hidden layer as 2, weight decay as 10$^{-4}$, and with a maximum number of iterations at 400 [51]. The R package pROC (version 1.14.0) was used for ROC analysis to compare the abilities of various models to distinguish between hyper- and hypo-methylated sites by the area under the curve (AUC) analysis [52].

t-SNE analysis, PCA analysis, and gene enrichment analysis
t-Distributed stochastic neighbor embedding (t-SNE) analysis was performed by R package t-sne (version 0.1-3) [53]. PCA was performed by R function princomp and visualized by first two principal components. KEGG and GO enrichment were analyzed online by DAVID 6.8 (https://david.ncifcrf.gov) [54, 55]. Ingenuity Pathway Analysis (IPA) was also used for enrichment analysis for more elaborate results with the $P$ value cutoff set at 0.05 [26].

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s13148-020-00851-3.

Additional file 1. Supplementary figures and tables

Acknowledgements
The authors gratefully acknowledge Dr. Steven J Schrodi, Dr. Emily A. Andreae and Dr. Ingrid Glirich from the Center for Precision Medicine Research (CPMR), Marshfield Clinic Research Institute (MCRRI), for reviewing, commenting on, and editing their manuscript.

Authors’ contributions
JF developed the analysis method and drafted the manuscript. JL recruited the patients, provided clinical expertise on colorectal adenomas, and interpreted the results. JF, HZ, and SG performed the data analysis. SG and DZ reviewed and edited the manuscript and provided respective scientific expertise in result interpretation. CT and YZ conducted the array experiments. WW collected and prepared tissue samples and collected the results of clinical assays. SD funded and supervised all clinical procedures for enrolled patients. DZ and CZ designed the study, supervised all experiments and analysis, provided molecular and cellular biology advice, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding
This study is funded by the Youth 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).

Availability of data and materials
All methylation array data are available at GEO under accession number GSE139404. Other public data involved in this study included GSE68060, GSE68838, GSE77954, GSE7965, GSE81211, GSE101764, GSE107352, GSE75546, GSE40279, and E-MTAB-6450.

Ethics approval and consent to participate
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). Informed written consent was obtained from all the patients and volunteers prior to the procedure.

Consent for publication
Not applicable

Competing interests
The authors disclose no potential competing interests.

Author details
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 Department of Medical Genetics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53726, USA. 4 Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA. 5 Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China. 6 University of Chinese Academy of Sciences, Beijing 100049, China.

Received: 21 January 2020 Accepted: 8 April 2020

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s13148-020-00851-3.

References
1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7–30.
2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu QX, He J. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115–32.
3. Kupers EJ, Grady WM, Lieberman D, Seufferlein T, Sung JJ, Boelens PG, van de Velde CJ, Watanabe T. Colorectal cancer. Nat Rev Dis Primers. 2015;1:15065.
4. 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. Nat Genet. 2017;49(4):435–42.

Supplementary figures and tables

Additional file 1. Supplementary figures and tables

Executive summary
The aim of the study was to develop and validate a non-invasive plasma DNA methylation signature that could distinguish high-grade adenomas (HGA) from normal tissue, with potential diagnostic and clinical implications.

Methods
A total of 518 samples from patients with colorectal adenomas were recruited from Peking University Third Hospital, Beijing, China. Plasma DNA was isolated and bisulfite sequenced to identify differentially methylated regions (DMRs) between high-grade adenomas (HGA) and normal tissue.

Results
A total of 1039 DMRs were identified, of which 354 were classified as hyper-methylated and 685 as hypo-methylated.

Discussion
The developed methylation signature showed high accuracy in distinguishing HGA from normal tissue, with potential implications for improving diagnostic and therapeutic approaches for colorectal adenomas.

Conclusions
The study highlights the potential of plasma DNA methylation as a non-invasive biomarker for the early detection and management of colorectal adenomas.
5. Wang X, Wang L, Guo S, Bao Y, Ma Y, Yan F, Xu K, Xu Z, Jin L, Lu D, 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. 2013;27(10):4194–203.

6. Guo S, Yan F, Xu J, Bao Y, Zhu J, Wang X, Wu J, Li Y, Pu W, Liu Y, et al. Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC). Clin Epigenetics. 2015;7:3.

7. Zhao Y, Xue F, Sun J, Guo S, Zhang H, Qiu B, Gong J, Gu J, Zhou X, Wang 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):30.

8. Patai AV, Molnár B, Kalmár A, Scholler A, Tóth K, Tulassy Z. Role of DNA methylation in colorectal carcinogenesis. Dig Dis. 2012;30(3):310–5.

9. Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. Gastroenterology. 2008;135(4):1079–99.

10. Hidaka H, Higashimoto K, Aoki S, Mishima H, Hayashida C, Maeda T, Koga Y, Yatsuji H, Joh-K, Nishiro H, et al. Comprehensive methylation analysis of imprinting-associated differentially methylated regions in colorectal cancer. Clin Epigenetics. 2018;10(1):150.

11. Shi YX, Wang Y, Li X, Zhang W, Zhou HH, Yin JY, Liu ZQ. Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer. BMC Genomics. 2017;18(1901).

12. Lindqvist BM, Wingren S, Motlagh PB, Nilsson TK. Whole genome DNA methylation signature of HER2-positive breast cancer. Epigenetics. 2014;9(8):1149–62.

13. Raggi C, Invernizzi P. Methylation and liver cancer. Clin Res Hepatol Gastroenterol. 2013;37(6):564–71.

14. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–92.

15. Morris MR, Latif F. The epigenetic landscape of renal cancer. Nat Rev Nephrol. 2013;9(7):417–20.

16. Herman JM, Merlo A, Mao L, Lapidus RG, Issa J-P, Davidson NE, Sidransky D, Baylin SB. Inactivation of the CDKN2/p16/MTS1 gene is frequently defective human tumor cell lines. Cancer Res. 1997;57(5):808.

17. Kane MF, Loda M, Gaida GM, Lipman J, Mishra R, Goldman H, Jessup JM, Herman JG, Merlo A, Mao L, Lapidus RG, Issa J-P, Davidson NE, Sidransky D, Baylin SB. Inactivation of the CDKN2/p16/MTS1 gene is frequently defective human tumor cell lines. Cancer Res. 1997;57(5):808.

18. Yuki Izumi, Shinsaku Kadowaki, Masahiro Inamura, Hiroshi Hatakeyama, and Masashi Arima. Regulation of the transcriptional activity of the Na+ channel by DNA methylation in embryonic stem cells. Genes Cells. 2009;14(4):405–15.

19. Yuki Izumi, Shinsaku Kadowaki, Masahiro Inamura, Hiroshi Hatakeyama, and Masashi Arima. Regulation of the transcriptional activity of the Na+ channel by DNA methylation in embryonic stem cells. Genes Cells. 2009;14(4):405–15.

20. Tóth K, Sipos F, Kalmár A, Patai AV, Wichmann B, Stoehr R, Golicher C, Schellerer V, Tulassy Z, Molnár B. Detection of methylated SEPT9 in plasma is a reliable screening method for both left- and right-sided colon cancers. PLoS One. 2012;7(9):60000.

21. Tóth K, Waskerrott R, Sipos F, Kalmár A, Wichmann B, Leiszer T, Valcz G, Juhász M, Miheller P, Patai Á, et al. Detection of methylated septin 9 in tissue and plasma of colorectal patients with neoplasia and the relationship to the amount of circulating cell-free DNA. PLoS One. 2014;9(9):e115415.

22. Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, Wang W, Sheng H, Pu H, Mo H, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Sci Transl Med. 2020;12(525).

23. Wang Q, Yan L, Hu Q, Sucheston LE, Higgins MJ, Johnson CS, Smigal DJ, Liu S, JIA: an R package for high-throughput analysis of Illumina’s 450K Infinium methylation data. Bioinformatics. 2012;28(5):729–30.

24. Qu X, Sandmann T, Frierson H Jr, Fu L, Fuentes E, Walter K, Ohrak K, Rumpel C, Moskaluk C, Lu S, et al. Integrated genomic analysis of colorectal cancer progression reveals activation of EGR4 through demethylation of the EGR4 promoter. Oncogene. 2016;35(50):6403–15.

25. consortium B. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. 2016;34(7):726–37.

26. Kang K, Bae JH, Han K, Kim ES, Kim TD, Yi J. A genome-wide methylation approach identifies a new hypermethylated gene panel in ulcerative colitis. Int J Mol Sci. 2017;18(16).

27. Barrow TM, Klett H, Töth R, Bohm J, Gigic B, Hubermann N, Schere D, Schrotz-King P, Skender S, Abbenhardt-Martín C, et al. Smoking is associated with hypermethylation of the APC 1A promoter in colorectal cancer: the ColoCare Study. J Pathol. 2017;243(3):366–75.

28. Damaso E, Castillejo A, Arias MDM, Canet-Hermida J, Navarro M, Del Valle J, Campos O, Fernandez A, Marin F, Turchetti D, et al. Primary constitutional MLH1 epimutations: a focal epigenetic event. Br J Cancer. 2018;118(9):976–87.

29. Bornmann F, Rodriguez-Paredes M, Laitschka F, Edelmann D, Musch T, Benner A, Bergmann Y, Dieter SM, Ball CR, Gilhm H, et al. Cell-origin DNA methylation signatures are maintained during colorectal carcinogenesis. Cell Rep. 2018;23(11):3407–18.

30. Anseeij H, Jaffe AE, Corradina Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(1):1363–9.
52. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
53. Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008;9(2):2579–605.
54. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1):44–57.
55. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.

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