March 2022

Cell-free DNA Methylation Signatures in Cancer Detection and Classification

Jinyong Huang
University of South Florida

Cell-free DNA Methylation Signatures in Cancer Detection and Classification

by

Jinyong Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Tumor Biology
College of Arts and Sciences
University of South Florida

Major Professor: Liang Wang, M.D., Ph.D.
Alvaro Monteiro, Ph.D.
Theresa Boyle, M.D., Ph.D.
Jong Park, Ph.D.
Mingxiang Teng, Ph.D.

Date of Approval:
March 4th, 2022

Keywords: Liquid biopsies, DNA methylation profiling, cfMBD-seq, Cancer biomarker

Dedication

This dissertation is dedicated to my parents (Weixiong Huang and Hong Lin) and my wife (Bidian Zheng). I owe my deepest gratitude to the support from my family.

Acknowledgments

I would like to express my deepest appreciation to my major professor, Liang Wang, for his tremendous support and guidance throughout the dissertation research. I am incredibly grateful to Alex C. Soupir, Jing Jia, Qianxia Li, Yijun Tian, Muosa M Almubarak, and other members of Dr. Wang's laboratory. I will never forget the invaluable suggestions in research and life offered by the lab members. I would also thank to our collaborators: Brandon J Manley, Bruna Pellini Ferreira, and Andrew Chang. I would like to extend my sincere thanks to my committee members: Alvaro Monteiro, Theresa Boyle, Jong Park, and Mingxiang Teng, for their generous advice and guidance. I would like to extend my appreciation to Ken Wright, Tiffany T. Ferrer, Danielle C Dorsett, and all other members of the Cancer Biology Program of Moffitt Cancer Center. I would thank to Sean J Yoder, Andrew T Smith, Lan M Zhang, and all other members of the Molecular Genomics Core. I would thank to Shawn C Brass, and Marek Wloch, and all other members of the Tissue Core. I would thank to Margaret Penichet, Gwen Mickens, and all other members of the Collaborative Data Services Core. I would thank to Qianxing Mo and all other members of the Biostatistics and Bioinformatics Core.

Table of Contents

List of Figures

List of Tables

Abstract

Early detection of cancer is believed as one of the best solutions to improve the therapeutic outcomes and overall survival of cancer patients. Analysis of circulating nucleic acids in bodily fluids, referred to as “liquid biopsies”, is rapidly gaining prominence for this purpose. Cell-free DNA (cfDNA) methylation has emerged as a promising biomarker for early cancer detection, tumor type classification, and treatment response monitoring. Currently, most cfDNA methylation profiling technologies are based on bisulfite conversion, while enrichment-based methods such as cfMeDIP-seq are beginning to show potential. To expand the use of enrichment-based methods in cfDNA methylation profiling, here, we report an ultra-low input method based on methyl-CpG binding proteins capture, termed cfMBD-seq. We optimized the conditions of cfMBD capture by adjusting the amount of MethylCap protein along with using methylated filler DNA. Our data showed high genome-wide correlation between cfMBD-seq with 1 ng input and the standard MBD-seq (>1000 ng input). Compared with the most commonly used HM450K assay, our results showed that cfMBD-seq reliably detected 94% of the methylated CpG islands detected by HM450K, while correctly classifying 98% of non-methylated sites (AUC=0.995). We also found that cfMBD-seq outperforms cfMeDIP-seq in the enrichment of high-CpG-density regions such as CpG islands, which play an important role in the regulation of normal biological functions and diseases. To identify the clinical feasibility of cfMBD-seq, we applied cfMBD-seq to profile the cfDNA methylome using plasma samples from colorectal (N=13), lung (N=12), pancreatic (N=12) cancer patients, and non-cancer controls (N=16). We identified 1759, 1783, and 1548 differentially hypermethylated CpG islands (DMCGIs) in lung, colorectal, and pancreatic cancer patients, respectively. Interestingly, the vast majority of DMCGIs were overlapped with aberrant methylation changes in the corresponding primary tumor tissues, indicating that DMCGIs detected by cfMBD-seq were mainly driven by tumor-specific DNA methylation patterns. From the overlapping DMCGIs, we carried out machine learning analyses and identified a set of discriminating methylation signatures that had robust performance in cancer detection and classification. Overall, our study demonstrates that cfMBD-seq is a powerful tool for sensitive detection of tumor-derived epigenomic signals in cfDNA. Our findings will help to expand on existing blood-based molecular diagnostic tests and identify novel methylation biomarkers for early cancer detection and classification.

Chapter 1: Overview of cell-free DNA methylation analysis (Literature review)

Parts of this section were previously published by Cancers, a peer-reviewed, open access journal of oncology, published semimonthly online by MDPI.

[1] Huang, J.; Wang, L. Cell-Free DNA Methylation Profiling Analysis - Technologies and Bioinformatics. Cancers (Basel) 2019, 11, doi:10.3390/cancers11111741.

1.1 Background

1.1.1 Early cancer detection

Cancer is one of the leading causes of death worldwide and the total number of diagnosed cancer cases keeps increasing globally [2]. The dismal mortality rates seen in patients with these malignancies are associated with advanced stage at the time of diagnosis. To improve the therapeutic outcomes and overall survival of this patient population, detection of cancer at an early stage is believed as one of the best solutions. Existing clinical interventions can be more effective for pre-invasive tumors before clinical symptoms appear. Medical interventions such as surgical resection are curative for most types of localized cancers that have not metastasized [2]. Successful examples of early cancer detection including mammography, Pap smear, colonoscopy/fecal test, and low-dose chest computed tomography have helped reduce the mortality of breast, cervical, colorectal, and lung cancer, respectively [3-6]. Despite these successes, recommendations for cancer screening in general populations continue to be debated due to the unacceptably high false positive rates of existing tests and potential overdiagnosis of nonlethal cancers. Prostate-specific antigen test, a widely used blood biomarker test for prostate cancer, is an example of the potential consequences of high false positives and unnecessary medical interventions [7,8]. As a result, most screening tests have limited ability to detect cancers in the general populations because they are only practical when they are used to test individuals who have a high risk of developing the screened cancer. To overcome these limitations, several efforts have been made towards the investigation of novel assays that enable early cancer detection with high accuracy. Among those, the use of liquid biopsies is rapidly gaining prominence for minimally invasive cancer detection and management [9-11].

1.1.2 Liquid biopsies

Tissue biopsies are the gold standards for the histological diagnosis and molecular characterization of cancers. However, these conventional sampling methods have shown limitations including the difficulty in obtaining biomaterial, sampling bias arising from tumor genetic heterogeneity, and even procedural complications [9]. Liquid biopsies, the minimally invasive sampling and analysis of analytes from blood or other body fluids, have emerged as a critical supplement to the tissue biopsies. Liquid biopsy analytes include circulating nucleic acids (cell-free DNA (cfDNA) as well as cell-free RNAs), circulating tumor cells (CTCs), extracellular vesicles, tumor-educated platelets, proteins, and metabolites. Since tumor-specific analytes can originate from different tissues, including metastatic tumor sites, liquid biopsies may represent a whole picture of a patient's malignancy and mitigate the problem of tumor heterogeneity [12,13]. Early on, liquid biopsies were focus on the genomic analyses such as somatic mutations and copy number alterations in CTCs or cfDNA [14]. Recently, more attentions have been paid to the transcriptome [15], the epigenome [16,17], the proteome [18], and the metabolome [19]. Moreover, novel artificial-intelligence-based bioinformatics methodologies are moving the liquid biopsy field towards multiparametric and multi-omic analyses [20]. Many studies have shown that liquid biopsies have the potential to provide information about primary tumors or metastases that are meaningful for early cancer detection, minimal residual disease monitoring, treatment selection, and response prediction [10,21,22].

1.1.3 Cell-free DNA

cfDNA in body fluids is a mixture of extracellular DNA fragments that are released from cells via apoptosis, necrosis, and active secretion [23]. The length of cfDNA is about 167 bp, corresponding to the unit size of a nucleosome (~147 bp) plus linker DNA associated with histone H1 [24,25]. cfDNA in the circulation has a short half-life between 16 minutes and 2.5 hours, enabling liquid biopsies as real-time and dynamic monitoring tools for the estimation of tumor burden [26]. A significantly higher level of cfDNA in cancer patients than in healthy individuals has been reported [26]. In addition, the increased cfDNA can decrease to a background level following surgery [27]. Results from these studies suggest that tumor cells-derived cfDNA is present in the blood of cancer patients. This tumor-derived cfDNA is called circulating tumor DNA (ctDNA). The concentration of ctDNA in plasma has been shown to correlate with tumor size and stage [28]. This association suggests the prognostic and predictive utility of ctDNA. Patients with detectable ctDNA have worse survival outcomes than those without [29,30]. Additionally, ctDNA has been found to be a significantly better prognostic predictor than commonly used tumor markers. Specifically, a higher level of ctDNA correlates with poorer clinical and radiological outcomes [31,32].

In addition to prognosis and prediction, several studies have demonstrated the potential of cfDNA in noninvasive early diagnosis of cancer. For example, mutations in cfDNA have been detected in plasma up to 2 years before cancer diagnosis [33]. However, genomic analysis of cfDNA in early-stage cancer is very challenging because cfDNA is often limited in yield and highly fragmented [34]. More importantly, ctDNA is extremely underrepresented in the high background of normal cfDNA [34]. The increasing availability and reliability of highly sensitive technologies, such as digital PCR (dPCR) and next-generation sequencing (NGS), are facilitating the detection of the trace amount of ctDNA in early-stage cancer [35]. Currently, cfDNA-based approaches that focus on the detection of cancer-associated single-nucleotide variants (SNVs) and somatic copy number variants (CNVs) have been applied into clinical settings [36]. However, SNV assays have limitations associated with confounding signals from blood cells due to clonal hematopoiesis [37]. Similarly, CNV assays are limited by minor differences between cases and controls resulting in a need for increased sequencing depths, which translates into higher costs [38]. More importantly, these genomic variations have not yet demonstrated robust tissue of origin classification across a broad range of tumor types. In contrast, given the inherited ability of tracing tissue of origin, cfDNA methylation has become a promising biomarker in liquid biopsies. Therefore, detection of tumor-specific cfDNA methylation signatures is believed to be a more robust approach for early cancer diagnosis.

1.1.4 DNA Methylation

Cancer is defined by not only extensive genetic changes but also additional biological processes such as the immune microenvironment and epigenetics [39]. Epigenetics include any process that guides genomic function and activity without altering the DNA sequence [40]. DNA methylation is a common epigenetic modification that plays an important role in eukaryotes that has important implications for normal biological functions and diseases. De novo DNA methylation is achieved by adding a methyl group to the fifth carbon of cytosine (5-methylcytosine, 5mC) via DNA methyltransferases (DNMTs) (**Figure 1**). This modification occurs most frequently at cytosine residues in the sequence context of 5'-C-phosphate-G-3' (CpG) [41]. When located at gene promoters, DNA methylation is a repressive mark of gene expression. DNA methylation can also occur at the gene bodies of actively transcribed genes, which enhances gene expression [42]. The current human genome build contains about 28 million CpGs, 60–80% of which are methylated [43]. In the mammalian genome, the majority of CpGs are methylated, except for CpG-rich regions called CpG islands (CGIs) and the nearby CpG shores (the region within 2 kb of the islands) [44]. On the contrary, the cancer methylome is characterized by global hypomethylation and CpG islands-specific hypermethylation [45,46]. Genome-wide DNA hypomethylation occurs predominantly at repetitive regions and causes chromosomal instability [47]. Hypermethylation in cancer cells is frequently observed in the transcriptional regulatory elements (promoters and enhancers) of tumor suppressor genes [48]. Beyond that, it has been reported that gene body DNA hypermethylation can activate the expression of oncogenes such as Homeobox genes [49]. These methylation aberrations can synergize with driver mutations to facilitate cancer development [50]. Growing evidence suggests that aberrant DNA methylation contributes to the tumorigenesis and tumor progression [51-54].

The detection of ctDNA methylation aberrations holds great promise as a blood-based test for cancer diagnosis for several reasons: First, aberrant DNA methylation occurs early during tumorigenesis and is abundantly present in the entire cancer process [55]. Second, in contrast to the highly heterogeneous nature of gene mutations, tumors of the same histological type tend to exhibit similar DNA methylation changes among different individuals [56]. Third, circulating components are shed from multiple body sites, while the methylation patterns of cfDNA are consistent with the tissues where they originated from [57]. All these advantages imply that cfDNA methylation may serve as feasible and reliable cancer biomarkers [58,59]. cfDNA methylation may be combined with traditional screening in primary diagnosis, choice of therapy, response to therapy prediction, minimal residual disease monitoring, and recurrence detection for better patient outcomes [60]. In this chapter, the technologies for DNA methylation analysis were summarized and their feasibilities for liquid biopsy applications were discussed. A brief overview of the bioinformatic approaches for the analysis of DNA methylation sequencing data were also provided. Overall, this chapter provides informative guidance for the selection of experimental and computational methods in cfDNA methylation-based studies.

1.2 Technologies for DNA methylation detection

1.2.1 Restriction enzyme-based methods

The use of methylation restriction enzymes (MREs) to cleave DNA at a specific nucleotide sequence is a classical method for methylation study. A pair of isoschizomers that recognizes the same sequence and cleavage point but exhibits different sensitivity toward the methylation state is used in these methods. Methylation-sensitive enzymes cleave only unmethylated DNA and leave the methylated DNA intact, while methylation-insensitive enzymes can cleave regardless the methylation status of the recognition sites (see Reference [61] for all available MREs). Based on this principle, array hybridization assays such as HpaII-tiny fragment enrichment by ligation-mediated PCR (HELP) [62], comprehensive high-throughput arrays for relative methylation (CHARM) [63], and methyl-sensitive cut counting (MSCC) [42] have been developed for DNA methylation analysis. More recently, MRE digestion has been coupled with sequencing technologies (MRE-seq) to study the role of DNA methylation in regulating alternative promoters [64,65]. After MRE digestion, the resulting DNA fragments are directly used for library preparation and sequencing. The sequencing results reveal the locations of the methylated CpG sites (undigested sites) within the enzyme recognition sequences and allows the estimation of relative DNA methylation levels. However, due to the limited CpG-containing recognition sites in intergenic and distal regulatory elements, MRE-seq tend to exhibit low coverage toward the whole methylome. Importantly, the severely fragmented nature of cfDNA restricts the application of

MRE-seq for cfDNA methylation profiling as some restriction sites may have been destroyed. To address this issue, methylated DNA sequencing (MeD-seq), which takes advantages of the LpnPI restriction enzyme, has been developed [66]. Unlike other methylation-sensitive enzymes, LpnPI has less specific recognition sites and its activity is limited by a short template size. Therefore, LpnPI is suitable for detecting DNA methylation in both CpG-dense and CpG-poor regions. MeD-seq has been applied for cfDNA methylation profiling and showed great potential in the discovery of novel methylation signatures and disease load monitoring [67].

1.2.2 Bisulfite conversion-based methods

Bisulfite conversion-based methods are ideal for the detection of DNA methylation because they provide qualitative and quantitative information for methylation sites at single base-pair resolution. Upon sodium bisulfite treatment on denatured DNA, unmethylated cytosine (C) residues are deaminated to uracil (U) and eventually converted to thymine (T) via DNA amplification, while methylated C residues remain unchanged during the conversion process (**Figure 1**) [68]. Analysis of bisulfite-converted DNA was previously coupled with Sanger sequencing and microarray for the investigation of specific DNA sequences. Nowadays, by integrating high-throughput next-generation sequencing (NGS), the entire methylome can be profiled in a single testing. However, bisulfite conversion causes substantial DNA degradation [69], which may result in loss of some critical information, especially for generally very trace amounts of cfDNA. Therefore, the library preparation and bisulfite treatment process should be optimized before implementation of bisulfite sequencing into cfDNA study. The key optimization is to perform end repair and methylated adapter ligation before bisulfite treatment, as this ensures the amplification of pre-ligated cfDNA (**Figure 2**). Other technical improvements include library preparation within a single tube and the employment of Dynabeads for purification and size-selection. The main bisulfite conversion-based technologies are summarized below.

1.2.2.1 Whole-genome bisulfite sequencing (WGBS)

WGBS presents as the most comprehensive and informative DNA methylation profiling technology [70]. The first human genome-wide, single-base-resolution DNA methylation profile was mapped by WGBS [71]. The major advantage of WGBS is that the methylation state of all cytosines, including low CpG density regions and non-CpG sites (CpA, CpT, and CpC), can be detected. However, since the whole methylome is targeted, WGBS is expensive when producing high depth data. To address the increasing demands for low DNA input, optimized methods such as single-cell bisulfite sequencing (scBS-seq) [72] and single-cell whole-genome bisulfite sequencing (scWGBS) [73] have been developed. scBS-seq adopts a post-bisulfite adapter tagging (PBAT) protocol to reduce bisulfite-induced DNA loss and eliminate the need for global amplification [74]. This highly efficient PCR-free method can generate library starting from 125 pg of DNA and is potentially applicable for cfDNA analyses [75]. On the other hand, scWGBS uses post-bisulfite single-stranded DNA library preparation [76]. WGBS has been attempted for mapping cancer-associated cfDNA methylation in metastatic breast cancer [77]. However, as the cost of large-scale WGBS is prohibitive, a sample pooling approach was adopted. As a result, a few prominent samples may overshadow the other samples and the complexity of the study is weakened. Fortunately, the costs of sequencing have continuously decreased in recent years, making WGBS more economically feasible. Most importantly, a large-scale WGBS study has been performed using cfDNA of 1493 cancer patients across more than fifty cancer types and 1135 non-cancer controls by the Circulating Cell-free Genome Atlas (CCGA) project [78]. More than

1000000 CpG sites of interest have been identified by this study and a target panel has been designed for further investigation using targeted bisulfite sequencing in other cohorts.

1.2.2.2 Reduced-representation bisulfite sequencing (RRBS)

To investigate the methylome more cost-effectively, RRBS was developed by integrating *Msp*I digestion, bisulfite conversion, and NGS [79,80]. This method preferentially enriches CpG-rich regions and can detect more than 83% of CGIs in mammalian genome [81]. To apply this method for limited cfDNA, single-cell RRBS (scRRBS) has been developed and the input is significantly decreased [82]. To avoid DNA loss, scRRBS integrates all the key RRBS reactions into a single-tube reaction so that DNA purification does not occur until the entire procedure is completed (**Figure 2**). This is achieved by modifying the buffer system and the reaction volumes to preserve the activities of different enzymes [83]. Capitalizing on these strategies, RRBS has been successfully used for methylation profiling of plasma cfDNA for the first time. Consequently, methylated haplotype analysis in plasma cfDNA has demonstrated the quantitative estimation of tumor load and tissue-of-origin mapping [84]. Methylation patterns identified by RRBS have shown 81.6% accuracy (49 out of 60 patients) in minimally invasive classification of pediatric solid tumors [85]. Like MRE-seq, RRBS has a relatively low coverage toward intergenic and distal regulatory elements because of the limited CpG-containing recognition sites.

1.2.2.3 Methylated CpG tandems amplification and sequencing (MCTA-seq)

MCTA-seq is a sensitive technique for detecting hypermethylated CGIs [86]. In this approach, a primer that consists of a semi-random sequence, a unique molecular identifier sequence, and an anchor sequence is used to amplify the bisulfite converted DNA at the 3'-end. Then, the methylated CpG tandem sites are selectively amplified using another primer containing the CpG tandem sequence CGCGCGG. Only fragments with methylated CpG can be further amplified and sequenced (**Figure 2**). This approach allows as little as 7.5 pg cfDNA input achieved by multiple rounds of amplification. Application of the MCTA-seq in cfDNA has identified dozens of DNA hypermethylation markers for effective detection of hepatocellular carcinoma [86], colorectal cancer [87], and gastric cancer [88]. MCTA-seq has also identified 146 tissue-specific hypermethylation markers that revealed the tissue of origin of cfDNA in liver and pancreas disease patients [89]. These biomarkers, including known and novel, demonstrated a high sensitivity and specificity for disease detection. However, since MCTA-seq targets CGCGCGG-rich CpG sites and preferentially enriches high CpG density regions, it can miss some important methylation signals in low CpG density regions.

1.2.2.4 Targeted bisulfite sequencing

Although allowing for the discovery of novel DNA methylation alterations, the methods discussed above is not practical in clinical settings, where a rapid turn-around time, cost-efficient methods and high depth of sequencing coverage are required [90]. A better way is to focus on specific regions of interest using target enrichment strategies. Consequently, targeted bisulfite sequencing is more clinically pragmatic because it is scalable, economical, and allows for a higher sequencing depth. Depending on target enrichment manners, targeted bisulfite sequencing may be categorized into the following two groups: amplicon-based enrichment and hybrid capture enrichment (**Figure 2**). The former uses specific PCR primers to amplify regions of interest after the bisulfite treatment, such as *EFC#93* primers for disseminated breast cancer [91], or *Vimentin* and *Fibulin 1* primers for hepatocellular carcinoma [92]. The latter is performed in solution using biotinylated oligos such as probes to capture complementary sequences from the bisulfite converted library. Namely, 5’-biotinylated capture probes are used to specifically pull-down DNA fragments that contain target CpG sites. A comprehensive methylation sequencing assay targeting 9223 consistently hypermethylated CpG sites in plasma cfDNA from 68 patients with advanced cancers and 66 non-cancer controls has successfully detected the presence of cancer and classified cancer type with high accuracy (AUC = 0.969; Sensitivity: 83.8%; Specificity:100%) [93]. Additionally, using a target panel of 103456 distinct regions (17.2 Mb, 1116720 CpG sites) identified from WGBS, the CCGA project assessed the performance of cfDNA methylation signatures in the detection and localization of multiple cancer types across all stages at high specificity (Accuracy: 93%; Sensitivity: 55.2%; Specificity: 99.3%) [78]. Although targeted bisulfite sequencing has been investigated to have high clinical value, this method is constrained by the relatively complicated primer and probe design procedures, especially for bisulfite-converted sites.

1.2.2.5 Methylation array

Illumina Infinium HumanMethylation450 BeadChip (HM450K) contains predesigned probes for more than 450k methylation sites that cover 96% of the CGIs [94] and dominated as the method of choice for the cancer methylome studies before the prevalence of NGS [95]. Infinium MethylationEPIC BeadChip (HM850K), a further developed version, covers more than 850k CpG methylation sites, including almost all sites on the HM450K plus additional CpG sites in the enhancer regions [96]. Currently, a huge number of HM450K datasets have been archived on the Gene Expression Omnibus (GEO) [97] and The Cancer Genome Atlas (TCGA) [98]. They have become outstanding public resources for the discovery of novel DNA methylation markers [99,100] and the validation of new established DNA methylation assays [101]. As for liquid biopsies, the Infinium methylation array has been applied for the discovery of methylation biomarkers for colorectal cancer [102], hepatocellular carcinoma [103] and prostate cancer [104]. The methylation data have also been used for the deconvolution of the plasma methylome for the inference of tissue origins of cfDNA [57]. However, all array-based methods have a drawback in poor genome-wide coverage of all CpG sites, resulting in the loss of methylation contexts.

1.2.2.6 Methylation-specific PCR (MSP)

MSP is based on the use of two distinct methylation-specific primer sets to detect the DNA of interest. The methylated primers can amplify both bisulfite-converted methylated DNA and untreated DNA, while the unmethylated primers are only specific for bisulfite-converted unmethylated DNA [105]. Taking advantage of real-time PCR, several quantitative MSP (qMSP) protocols [106-108] and methylation-sensitive high-resolution melting analysis (MS-HRM) [109] has been developed for DNA methylation analyses. These technologies have been widely used in the identification and validation of aberrant DNA methylation in cfDNA in breast cancer [110-112], pancreatic cancer [113], ovarian cancer [114], cholangiocarcinoma [115], hepatocellular carcinoma [116,117], prostate cancer [118], and lung cancer [119,120]. With the emergence of dPCR, digital methylation-specific PCR (dMSP) has also been developed for the screening and validation of cfDNA-based methylation biomarkers for breast cancer [121], prostate cancer [122], and colorectal cancer [123-125]. Yet, these individual markers only provided a limited picture of the whole tumor methylome. Therefore, a combination of multiple markers is highly recommended in the clinical settings to solve the problem of tumor heterogeneity and to ensure a high sensitivity and specificity.

1.2.2.7 Oxidative bisulfite conversion

Ten-eleven translocation (TET) enzymes can oxidize 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), which is known as DNA demethylation (**Figure 1**) [126,127]. Taking advantage of the fact that cytosines in 5fC and 5caC are not protected from deamination by sodium bisulfite (**Figure 1**), oxidative bisulfite sequencing (OxBS-seq) [128] and TET-assisted bisulfite sequencing (TAB-seq) [129] have been developed, respectively.

However, as longer bisulfite treatment and oxidative environment are needed for the efficient conversion of 5mC to 5fC or 5caC, more DNA degradation and DNA damage may occur [130]. Application of these methods in liquid biopsies need further investigation.

1.2.3 Enrichment-based methods

Enrichment-based methods are based on the use of antibodies or proteins to pull-down the methylated genomic regions for subsequent analysis, while the unmethylated fractions are excluded by stringent washing. Compared to WGBS, enrichment-based methods have not only shown a similar sensitivity and specificity [131], but also have many other advantages. They are cost-effective because only the enriched fragments are sequenced, therefore, many indexed samples can be pooled simultaneously for a single NGS run. Unlike bisulfite conversion-based methods, the enrichment approaches do not involve cytosine conversion, therefore keep DNA sequences intact. Furthermore, they can discriminate 5mC from 5hmC due to the protein-binding specificity. However, these methods have a relatively low resolution (100-300 bp), and therefore could not discriminate the exact methylation state of a single CpG site. Additionally, these methods tend to exhibit biases toward hypermethylated regions. As the standard protocol of these methods require relatively large amount of DNA input, further optimizations in the library preparation and methylation enrichment are needed to apply them in cfDNA-based studies.

1.2.3.1 Methylated DNA immunoprecipitation sequencing (MeDIP-seq)

MeDIP was originally developed as an approach for the immunoprecipitation of methylated DNA followed by a microarray analysis [132]. A low DNA input protocol has been reported to reduce the required input from 5000 ng to 50 ng DNA. However, using less than 50 ng DNA as an input was not recommended due to insufficient methylation enrichment [133]. To apply MeDIP-seq for low-input cfDNA in liquid biopsies, cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) has been developed, where exogenous lambda DNA is used as a filler to increase the initial DNA input (**Figure 3**) [134]. The filler DNA ensures a constant antibody/DNA ratio and helps maintain a similar immunoprecipitation efficiency across different samples with different cfDNA yields, while minimizing non-specific binding and DNA loss [135]. With the help of filler DNA, the starting cfDNA can be reduced to 1-10 ng. Because the lambda DNA does not have sequencing adapters, and hence no subsequent amplification, the use of filler DNA would not interfere with the analysis of sequencing data. cfMeDIP-seq has shown high accuracy in the classification of a wide variety of cancer types [126]. cfMeDIP-seq has been applied in plasma cfDNA in renal cell carcinoma patients and the identified classifier performed accurate classification of patients across all stages [136]. Compared to cfDNA variant analysis, cfMeDIP-seq is significantly more sensitive for the detection of renal cell carcinoma [137]. The methylome profiled by cfMeDIP-seq revealed highly specific methylation signatures to detect and accurately discriminate common primary intracranial tumors that share cell-of-origin lineages [138].

1.2.3.2 Methyl-CpG binding domain protein capture sequencing (MBD-seq)

Instead of immunoprecipitation, the methyl-CpG binding domain (MBD) of methyl-CpG binding proteins (MBD2 or MECP2) can be used to pull down methylated DNA fragments with the help of magnetic beads [139]. It has been shown that MBD-based enrichment outperforms MeDIP in regions with a higher CpG density and identifies the greatest proportion of CGIs [140]. Therefore, integrating MBD-seq with liquid biopsies may facilitate the discovery of CGIs-specific hypermethylation signatures. A study has described a low DNA input MBD-seq protocol by adjusting the DNA to beads ratio and using more incubation time and more stringent wash conditions [141]. Using this protocol, MBD-seq with a 15 ng DNA input detected 93% of the methylated loci that were reliably detected using WGBS (sensitivity) at similar levels of the false positive rate (specificity). Even with as little as 5 ng DNA, MBD-seq had a 90% of sensitivity and equal levels of specificity relative to WGBS [141]. Therefore, this low-input technology may be suitable for liquid biopsy studies. Also, it is expected that the use of exogenous DNA as a filler to increase the initial input might increase the capture efficiency of MBD proteins as it does for immunoprecipitation.

1.2.3.3 5-hydroxymethylation profiling

Emerging evidence indicates that 5hmC not only acts as a relatively stable epigenetic marker in mammals [142] but also correlate with tumorigenesis and tumor progression [143]. Previously, studies have shown the reduced global 5hmC levels but increased regional 5hmC levels in various cancer tissues [144]. These observations suggest that 5hmC signatures may also be promising biomarkers for cancer diagnosis and prognosis. To profile hydroxymethylation in the trace amount of cfDNA, 5hmC-Seal (aka hMe-Seal) has been developed [145]. In 5hmC-Seal, an azide-modified glucose is first introduced by β-glucosyltransferase (β-GT) and subsequently biotinylated via click chemistry in selective chemical labeling (**Figure 3**). The biotinylated 5hmC is then enriched using streptavidin beads followed by NGS to determine the genomic distribution of 5hmC, where spike-in probes are adopted to test the 5hmC capture efficiency during the 5hmC-Seal assay. The proof-of-principle global analysis of hydroxymethylome in cfDNA has been reported [146]. Since then, the 5hmC-Seal has been used to identify a genome-wide pattern of cancer-associated 5hmC changes and tissue origins of such changes in plasma cfDNA from a patient-derived xenograft mouse model [147]. The method has also been used to detect aberrant 5hmC alternations in both gene bodies and promoter regions for non-small-cell Lung Cancer [148], colorectal cancer [149], hepatocellular carcinoma [150], and esophageal cancer [151].

1.3 Bioinformatics analyses of DNA methylation sequencing data

The general workflow for the bioinformatics analysis of DNA methylation sequencing data includes data processing and quality control, data visualization, statistical analysis (identification of differential methylation between different groups), and data interpretation. Below, the analysis strategies for DNA methylation sequencing data are provided. All these strategies are highly compatible when cfDNA is used as a starting material.

1.3.1 Quality controls and alignment

Before alignment, it is highly recommended to perform some quality control checks to ensure that the raw data is in high quality. FastQC can provide a QC report which can spot problems that originate either from the starting library material or from the sequencer [152]. Trimming tools such as Trimmomatic [153], Trim Galore [154] can be used to trimmed off the low-quality base calls and sequencing adapter from the 3' end of the reads. For faster data processing, fastp, an all-in-one FASTQ preprocessor, is recommended. fastp can perform quality control, adapter trimming, quality filtering, and per-read quality pruning with a single scan of the FASTQ data and it is 2–5 times faster than other FASTQ preprocessing tools [155].

Alignment of bisulfite sequencing data is challenging because the bisulfite converted DNA does not align to the reference genome. To address this issue, two algorithms have been developed: wild card algorithm (BSMAP [156], RRBSMAP [157], Pash [158], and GSNAP [159]) and three-letter algorithm (Bismark [160], BS Seeker 3 [161], and BRAT-BW [162]). The wild card algorithm allows both Cs and Ts in reads to map into Cs in the reference genome, while the three-letter algorithm converts all Cs in the reference genome and the reads into Ts, and thus standard aligners can be adopted [163]. Post-alignment quality control is important for bisulfite sequencing data to reliably quantify read counts and the methylation level per base. For example, base-calling quality should be checked because miscalled bases can be counted as C-T conversions. Since the end repair step in library preparation may introduce either methylated or unmethylated Cs [164], low quality bases on sequence ends should be trimmed to minimize false C-T conversions. It is also critical to check the unique alignment rates and insert lengths after trimming because bisulfite treatment can cause substantial DNA degradation. Incomplete bisulfite conversion may cause false positive results as unconverted unmethylated Cs are considered as methylated. To address this issue, spike-in sequences are usually added to measure the bisulfite conversion rate. As the majority of CpGs with high inter-population differences contain common single nucleotide polymorphisms (SNPs) [165], filtering out known C/T SNPs is highly recommended. Additionally, removal of duplicate reads that align to the same genomic position arising from PCR amplification should be considered. However, this is problematic in RRBS because by design reads start at the same position even if they are not PCR duplicates. Instead, one can remove regions with unusually high read counts.

Differing from bisulfite-based methods, standard aligner (bowtie2 [166], BWA [167]) can be directly used for the alignment of enrichment-based sequencing data, because no mutation is introduced during library preparation. In the cases of enrichment-based sequencing data, none of the post-alignment quality control issues regarding C-T conversions mentioned above need to be considered. Duplicate reads are increasingly likely to occur because reads are expected to align to a smaller methylation-enriched genome, where some duplicate reads occur by chance owing to the methylation enrichment, not the over-amplification. Poisson statistic has been used to determine the maximal number of duplicate reads allowed per genomic position [168].

1.3.2 DNA methylation calling

After a series of post-alignment quality control, the methylation level (a number ranging from 0 to 1) of individual CpG site can be calculated in the bisulfite sequencing data. This is simply done by counting the number of C-T conversions and dividing the number of Cs by the sum of Cs and Ts for each C. As a relative ratio, the methylation level would normalize the coverage difference at each CpG site, which vary dramatically due to genomic feature and amplification differences. Thus, additional normalization across different samples is not needed for bisulfite sequencing data. Moreover, the methylation level can be easily fit into many commonly used statistics models for differential methylation analysis as it is a continuous variable. However, the methylation level calculated from the CpG site with a low sequencing depth is less reliable.

Unlike bisulfite sequencing data, data from enrichment-based methods are usually analyzed by comparing the relative abundance of fragments. Generally, the genome is divided into non-overlapping adjacent windows of a specified width, and the number of reads in each window across all samples can be used for further analysis. For analysis involving multiple samples, data normalization is crucial to remove biases between samples or different batches. TPM (transcripts per million) and RPKM/FPKM (reads/fragments per kilobase of transcript per million reads mapped) are popular choices as they rescale read counts for differences in both the sequencing depth and fragment length [169]. However, TPM and RPKM/FPKM normalized data are not suitable for differential methylation analysis. Because the difference in biological methylation composition is not considered and the few hypermethylated regions between samples can skew the normalization. Instead, the trimmed mean of M-values (TMM) [170] and DESeq2 [171] normalization are highly recommended because they not only normalize the sequencing depth but also account for the methylation composition. They calculate a scaling factor via different algorithms and then read counts are normalized by the scaling factor. The genomic binning function and TMM normalization are implemented in the R package MEDIPS [168], while DESeq2 normalization is implemented in the R package DESeq2. More recently, a Bayesian statistical model that transforms the methylation enrichment read counts to absolute methylation levels has been developed and is implemented in the R package QSEA [172,173]. Additionally, a non-parametric method that uses isolated CpGs to estimate sample-specific fragment size distributions for the estimation of the methylation level of each CpG site has also been developed and is implemented in the R package RAMWAS [174,175].

While normalization is essential before differential expression analyses, visualization of data is also necessary whenever read counts between samples are compared. UCSC Genome Browser [176] and Integrative Genomics Viewer (IGV) [177] are usually used for data visualization. Methylation plotter [178] and Web Service for Bisulfite Sequencing Data Analysis (WBSA) [179] are specially designed for visualization DNA methylation sequencing data.

1.3.3 Determination of differential methylation

Following methylation calling, statistical tests can be employed to identify differential methylation between cases and controls. Differential methylation in cancer means CpG sites or regions that have different DNA methylation patterns between cancer patients and healthy individuals. For a comparative analysis of the methylation level (from 0 to 1) on a CpG site or region between multiple samples, statistical methods such as t-test, ANOVA, and nonparametric test (Mann–Whitney U test and Kruskal-Wallis test) may be used. T-test is used to examine if two sets of samples have significantly different means (R package BSmooth [180] and Minfi [181]). It is a parametric test and requires the samples to be independent and normally distributed. ANOVA is based on linear models for a multiple-group comparison (R package Minfi). Thus, it is more flexible to incorporate multiple clinical covariates and to accommodate different study designs. Considering the distribution of methylation level among the study population is unknown, nonparametric test is a safer approach in methylation studies as it is free of distribution assumption

(R package limma [182]). The Mann–Whitney U test evaluates if two comparison samples have identical medians and thus it is less affected by outliers. However, nonparametric test has less power and will be problematic when the sample size is small.

Bisulfite sequencing data can also be analyzed based on read counts. Read count is the number of methylated and unmethylated cytosines at a CpG site. Contingency table test (Fisher's exact or chi-square test), clustered data analysis [183], logistic regression, and the beta-binomial model [184-186] may be used for read counts based analyses. Contingency table test is the simplest method when replicates are not available, however, it does not take the variability of interindividual variation into consideration (R package COHCAP [187]). Clustered data analysis is an optimized version of the contingency table test as it incorporates the between-subject variability. This method can be considered as performing the chi-square test for independence in a series of contingency tables. However, large sample size is needed since the test is based on approximation. Logistic regression is a form of generalized linear model (GLM) for binary data (R package methylKit [188]). However, there are clearly more biological variability than the binomial assumption (only methylated and unmethylated) in the data. Regression methods allow adding covariates, such as age and sex into the tests, which are shown to be influential on the methylation level [189]. Among these models, the beta-binomial model is the best method for balanced sensitivity and specificity in differentially methylated cytosines (DMCs) detection (R package methylSig [185]) [190]. For enrichment-based read count data, well established RNA-seq data analysis R packages such as DESeq2 [191] and edgeR [192] can be directly applied to the normalized read count matrix for differential methylation analysis.

As the methylation level between neighboring CpG sites are potentially positively correlated, a combination of multiple adjacent CpG sites into a defined region called differentially methylated regions (DMRs) can reduce the number of hypothesis tests and thereby improve the statistical power [164]. DMRs can be determined by clustering nearby DMCs. DMRs can also be defined based on predefined regions, such as gene promoters and CpG islands, or adjacent CpG sites within user-defined non-overlapping windows across the whole genome [193]. To better measure weak methylation differences, increasing the biological replicates and sequencing depth present a good strategy to obtain more robust *p*-values. The inherent limitation of high dimensional data is false positive. Therefore, statistical results must be subjected to multiple testing corrections. Among all options, Bonferroni and false discovery rate (FDR) are the most commonly used multiple testing correction methods. Comprehensive evaluation of almost all tools and statistical methods for identifying DMRs for DNA methylation sequencing data has been summarized [190,194-198]. After the identification of DMRs, the regions of interest often need to be integrated with genome annotation datasets, which allows for determining whether the DMRs are related to genes and gene regulatory regions. The R package Genomation [199] and annotatr [200] are good annotation tools.

1.3.4 Identification of tumor-specific methylation profile

Due to the high background of normal cells-derived cfDNA, the ctDNA concentration in cfDNA is generally low in cancer patients. Therefore, it is challenging to identify ctDNA methylation, especially in early-stage cancer. One common strategy is to use the methylation profile of tumor-free peripheral blood mononuclear cells (PBMCs) as a negative control. By comparing DMRs between cancer cfDNA and healthy cfDNA to DMRs between cancer cfDNA and PBMC genomic DNA, the shared regions are considered to be tumor specific DMRs. This strategy has been applied in many studies that identified the ctDNA methylation signatures [99,134,146,147]. Additionally, a reference-based deconvolution algorithm has been developed for correcting cell-type heterogeneity [201]. This algorithm allows for the recovery of the original signal from a mixture of signal sources by using reference datasets. Therefore, it is suitable for the deconvolution of data from heterogeneous samples like cfDNA, using public available tissue-specific or cancer type-specific methylation data. The deconvolution algorithm has been successfully applied to estimate ctDNA content and differentiate tissue-of-origin in cfDNA of patients with lung or colorectal cancer [84]. Recently, probabilistic models have been formulated to identify ctDNA methylation. CancerLocator, a tool for non-invasive cancer diagnosis and tissue-of-origin prediction, is based on such a model [202]. By using Infinium HM450K data from TCGA, CancerLocator identified many CpG cluster features that have significant methylation variation across cancer and normal samples, as well as modeled methylation levels in different cancer types. Thus, the ctDNA burden and the likelihood of the presence of a specific cancer type can be inferred based on the informative CpG clusters. A further developed version, CancerDetector, adopts the joint methylation states of multiple adjacent CpG sites on an individual sequencing read and jointly deconvolutes the tumor fraction across all markers, has achieved a high sensitivity and specificity in detecting ctDNA methylation [203]. The Bayesian hierarchical model and methylation haplotype analysis share a similar strategy that enables information sharing across a cluster of neighboring CpG sites in order to enhance the statistical power [84,172,184].

1.4 Current challenges and future directions

In the early years, whole blood (buffy coat) DNA was preferentially used as a starting material for liquid biopsies. However, the high background of the hematopoietic cell genome may cause a false positive detection of tumor specific DMRs. Later, plasma was proven to be a superior source of cfDNA owing to the lower background levels of normal cfDNA [204]. Although superior, the cfDNA yield and ctDNA fraction are still limited in early-stage cancer plasma samples. To ensure successful and solid results, most biomarker studies are limited to plasma samples from metastatic and late-stage cancer patients. However, biomarkers identified from advanced-stage disease are not fully applicable for early-stage cancer due to the dramatic methylation changes during disease progression. That's why useful methylation signatures have yet not been fully established in clinical practice although cfDNA methylation profile in cancers has been known for a decade. One should be aware that if a certain region of DNA is not present in the sample, no target enrichment technique can retrieve it. Given the extremely low ctDNA content, the early-stage cancer plasma is expected to contain a low copy number of tumor genome equivalents. Therefore, the region with low sequencing read counts should be considered as a potential tumor-derived signal. Unfortunately, most analytical pipelines do not take such a scenario into consideration and filter out low-depth regions as part of routine quality control. Although most existing computational data analysis methods are applicable for cfDNA methylation sequencing data, further improvements in programing are still needed to detect the trace amount of tumor-derived methylation signal in early-stage cancer.

The use of NGS in epigenetic studies has significantly facilitated the discovery of DNA methylation biomarkers. However, these studies also face some challenges, including the lack of a uniform pipeline for both experimental and computational methods. Thus, different laboratories may identify different set of biomarkers even from the same type of disease or the same set of samples. In some cases, different researchers may have different interpretations on the same datasets. To eliminate the inconsistency arising from tumor heterogeneity and distinctive technical and analytical methodologies, integration of different methylation assays may be considered as a strategy of choice. It is believed that the combination of various methylation assays will guarantee the generation of more reliable biomarkers. If a genomic region can be identified by different methylation assays in different patient cohorts, it will be believed as a robust biomarker. The reliability of a novel detection assay can also be validated by existing assays. For example, the high consistency of methylation profile generated by cfMeDIP-seq was validated by comparing with traditional MeDIP-seq, RRBS, and WGBS [134]. Additionally, the combination of multiple methylation signatures will help to achieve a higher accuracy performance in a diagnostic or prognostic model. For instance, recent studies for the early detection of breast cancer and for the monitoring of treatment response in colorectal cancer used multiple methylation signatures to improve model prediction outcome [121,205]. Furthermore, the integration of cfDNA methylation analysis with other aberrant cfDNA alternations assays, such as copy number variations and point mutations, will also improve the diagnostic sensitivity and specificity.

Although promising, the integrated strategy will produce more complicated data, which requires a more sophisticated analytical algorithm. With the rapid development of new computational technologies, the use of machine learning for diagnostic and therapeutic decision making is receiving more popularity. For example, artificial intelligence systems have been adopted for methylation analysis in recent studies [110,134,206]. It is expected that machine learning will allow for the identification of trends and cancer-specific patterns with ease. However, machine learning has many obstacles. First, machine learning analysis requires massive training data sets. However, studies aimed to prove the robustness of cfDNA biomarkers often possess inadequate sample size and statistical competency. Second, cohort information such as sex, age, cancer type and stage, as well as diverse preanalytical factors (sample collection and storage) are necessary. However, it is always difficult to collect comprehensive clinical information for all patients. Third, machine learning analysis is a complex and time-consuming process. Therefore, adept computer programing skill as well as statistical knowledge are required for scientists to comprehensively analyze and interpret the vast amount of data. Significant efforts are still needed to fully apply cfDNA methylation signatures for cancer detection and outcome prediction in the clinical settings. Besides cfDNA methylation, other epigenetic biomarkers have also been explored for liquid biopsies. With the rapid development of cell sorting technologies, investigation of the methylome in circulating tumor cells (CTCs) has become possible [207]. A fundamental connection between phenotypic features of CTCs and DNA methylation dynamics in stemness and metastasis has been identified recently [208]. More knowledge regarding the CTC methylome remains to be further explored. Meanwhile, cell-free RNA methylation and cfDNA fragmentation patterns also deserve more research attentions in the future [209,210].

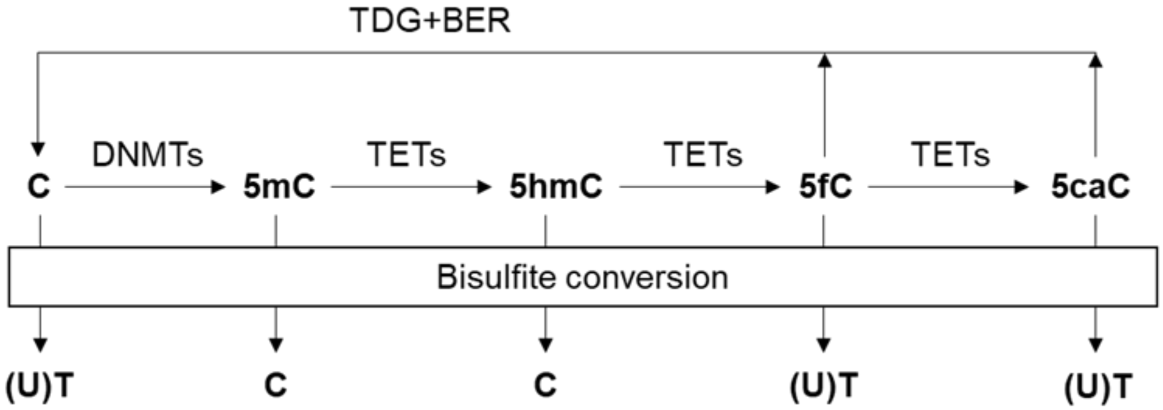


Figure 1. The dynamic regulation of DNA methylation

Cytosine variants and their products by bisulfite conversion. DNA methyltransferases (DNMTs) convert unmodified cytosine (C) to 5-methylcytosine (5mC) by adding a methyl group. Ten-eleven translocation (TET) enzymes oxidize 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Thymine DNA glycosylase (TDG) and the base excision repair (BER) pathway allow for regeneration of C from 5fC and 5caC. Upon bisulfite treatment, unmethylated cytosine (C) is deaminated to uracil (U) and eventually converted to thymine (T) via DNA amplification, while methylated C remains unaffected. 5hmC also protects C from deamination, while 5fC and 5caC do not.

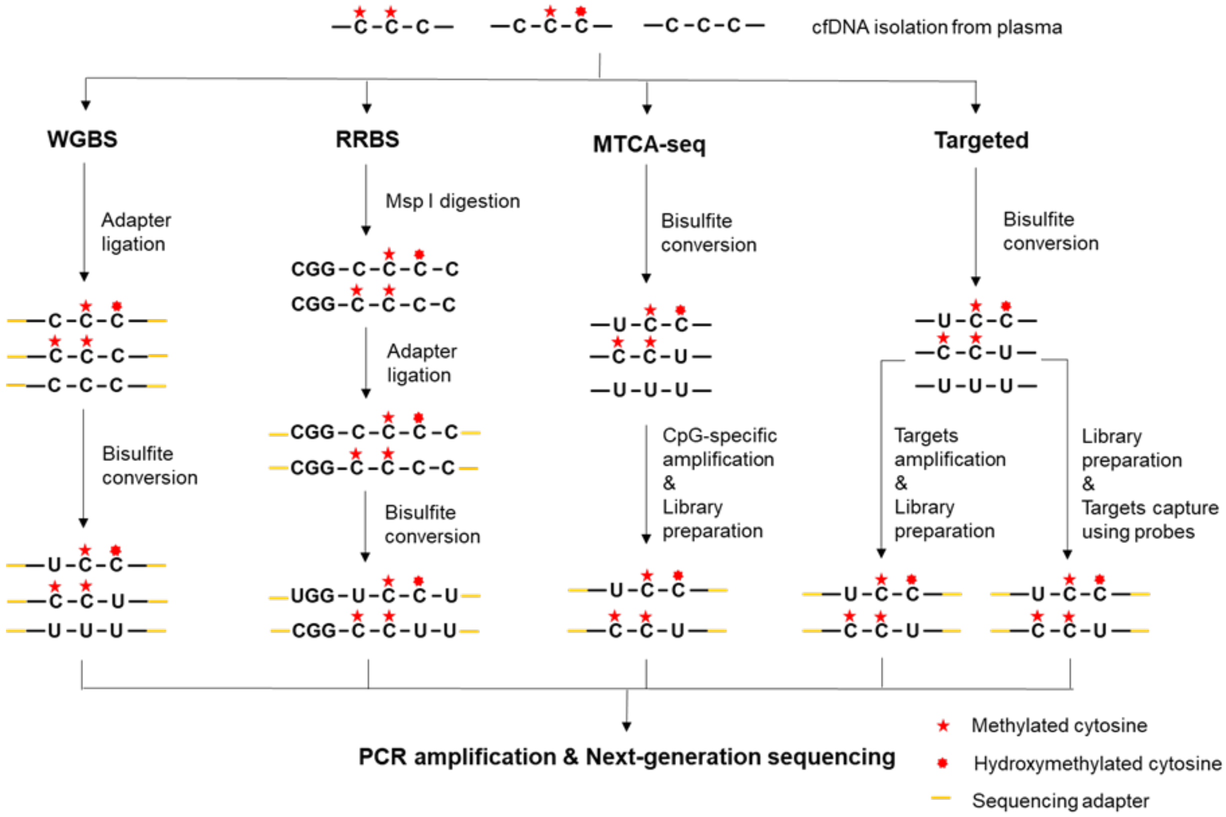


Figure 2. Schematic diagram of bisulfite-based methylation profiling technologies

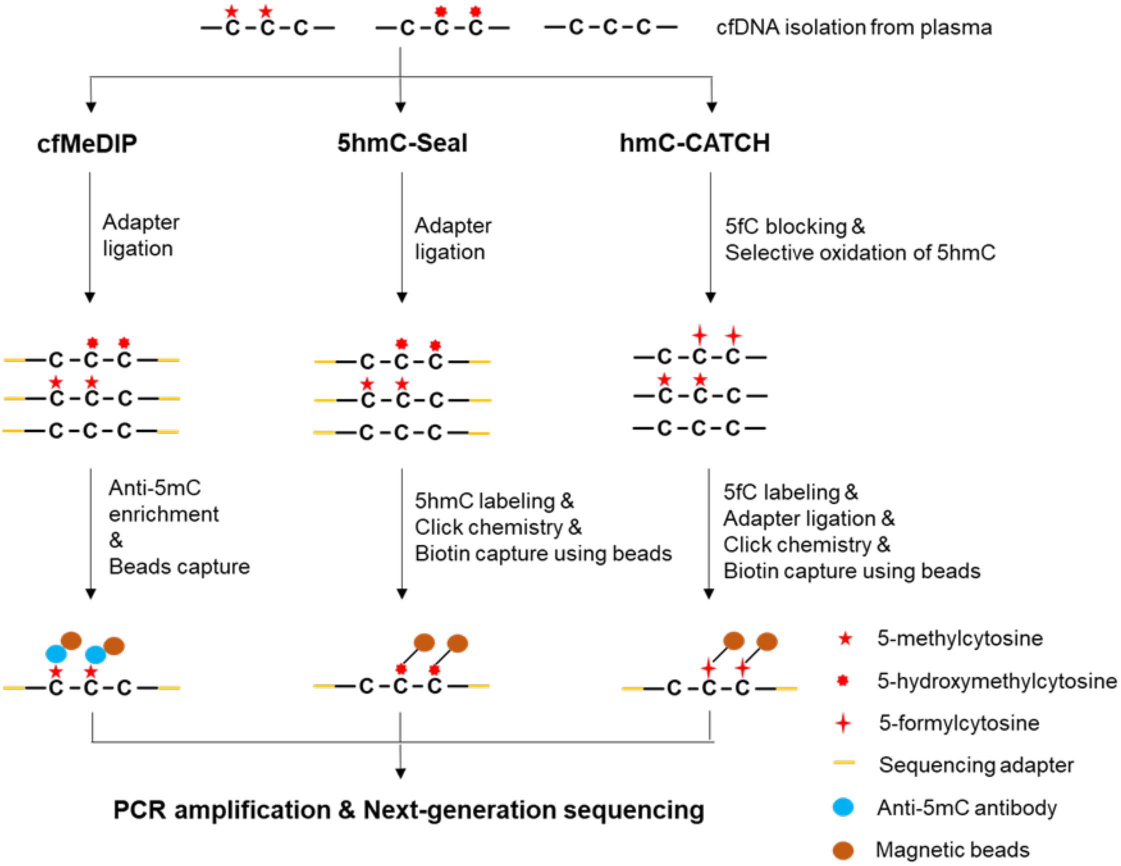


Figure 3. Schematic diagram of enrichment-based methylation profiling technologies

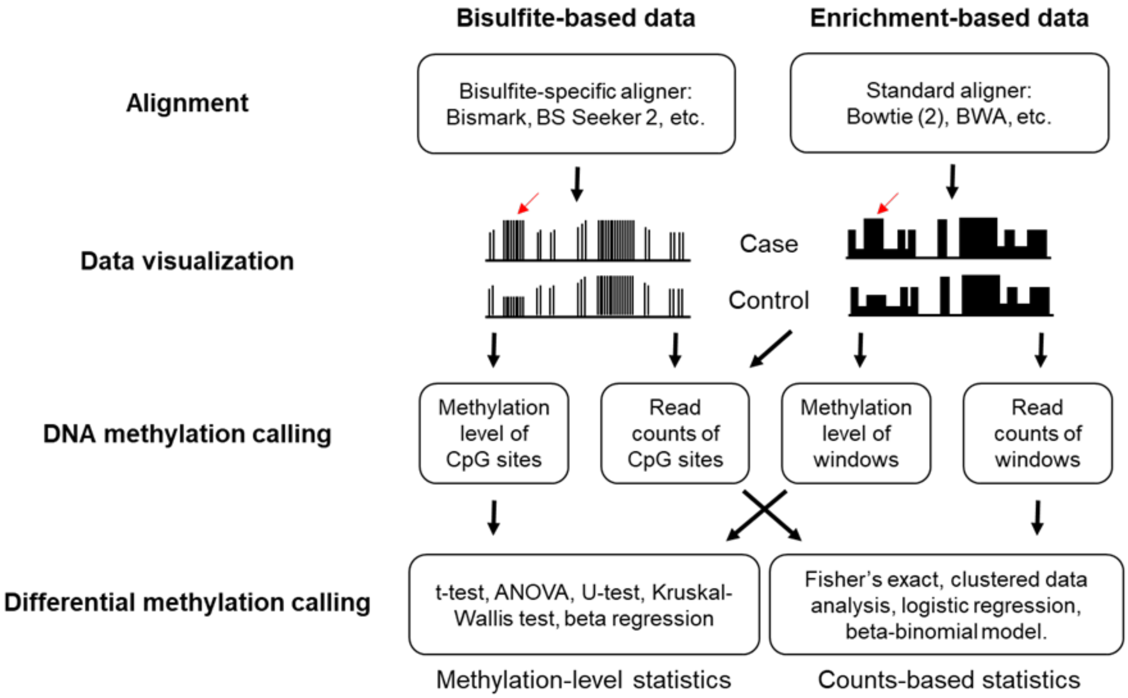

Bisulfite-based data
Enrichment-based data
Alignment
Bisulfite-specific aligner: Bismark, BS Seeker 2, etc.
Standard aligner: Bowtie (2), BWA, etc.
Data visualization
Case
Control
DNA methylation calling
Methylation level of CpG sites
Read counts of CpG sites
Methylation level of windows
Read counts of windows
Differential methylation calling
t-test, ANOVA, U-test, Kruskal-Wallis test, beta regression
Fisher's exact, clustered data analysis, logistic regression, beta-binomial model.
Methylation-level statistics
Counts-based statistics

Figure 4. Computational pipeline for DNA methylation sequencing data analysis

Table 1. Strengths and weaknesses of all major methylation assays for liquid biopsies

Class	Technology	Strength	Weakness	Cost
Restriction enzyme-based		-High CGI coverage	-Low resolution -Limited to regions in proximity to restriction enzyme sites	
	qPCR or dPCR	-Allows ultra-low DNA input -Easy primer design	-Loci-specific studies only	Low
Bisulfite-based		-Single-based resolution	-Substantial DNA degradation during bisulfite treatment -Cannot discriminate between 5mC and 5hmC	
	WGBS	-The most comprehensive profiling of the whole methylome	-Relatively low sequencing depth	High
	RRBS	-High CGIs coverage	-Limited to regions in proximity to restriction enzyme sites	Moderate
	MTCA-seq	-High CGIs coverage	-Limited to CGIs and might decrease other methylation backgrounds	Moderate
	Targeted	-Detect target CpG sites at high coverage	-Complicated primer or probe design	Low
	Microarray	-Pre-designed panel covering hotspot methylation	-Low genome-wide coverage of CpGs	Low
	qMSP or ddMSP	-Allows ultra-low DNA input	-Loci-specific studies only -Complicated primer or probe design	Low
Enrichment-based		-No mutation introduced	-Low resolution -Biased toward hypermethylated regions	
	MeDIP-seq	-Antibody is specific to 5mC	-Less sensitive in regions with high CpG density than MBD-seq	Moderate
5hmC profiling		-Specific to 5hmC	-High sequencing depth is required as 5hmC has a low abundance	
	5hmC-Seal	-Ensures accurate capture of DNA containing 5hmC	-Low resolution	Moderate
	hmC-CATCH	-Single-based resolution	-Oxidative environment would cause DNA damage	Moderate

Chapter 2: Optimization of MBD-seq for low input cell free DNA methylome profiling

Parts of this section have been accepted for publication in Epigenetics, published by Taylor & Francis.

[101] Huang, J.; Soupir, A.C.; Wang, L. Cell-free DNA methylome profiling by MBD-seq with ultra-low input. Epigenetics 2021, 1-14, doi:10.1080/15592294.2021.1896984.

2.1 Introduction

Currently, the majority of cell-free DNA (cfDNA) methylation profiling technologies are based on chemical treatment of the DNA with bisulfite [1]. Although whole-genome bisulfite sequencing (WGBS) of cfDNA has been attempted, this approach is not feasible for most clinical studies because of prohibitive cost and limited information recovery owing to the low genome-wide abundance of CpGs [211,212]. To address this issue, highly sensitive targeted assays such as targeted bisulfite sequencing and digital methylation-specific PCR have been developed [78,213]. The scalable and economical targeted bisulfite sequencing has been applied in large-scale cfDNA-based clinical studies [78,93]. High accuracy diagnostic prediction models of hepatocellular carcinoma and colorectal cancer have been established from a large cohort of patients and healthy controls [214,215]. However, the target methylation markers of these studies were selected from Infinium HumanMethylation450 BeadChip (HM450K) data. This methylation array is known to have selection bias and poor genome-wide coverage of all methylation sites, which may result in omission of important targets [99]. Similarly, the applications of quantitative and digital methylation-specific PCR are restricted by their low throughput nature. Alternatively, enrichment-based methylation profiling methods have shown a similar sensitivity and specificity when compared to bisulfite conversion-based methods [131]. Methylated DNA can be captured by methyl-CpG binding proteins (MBD) or anti-5mC antibodies (MeDIP) that have high affinity toward methylated CpGs. One critical limitation of such methods for liquid biopsy applications is that a relatively large amount of input DNA (ideally >1000 ng) is required, while the yield of cfDNA is typically low (2~10 ng/ml plasma). To address this issue, Shen et al. optimized the MeDIP-seq protocol to allow methylome analysis of small quantities (1–10 ng) of cfDNA, termed cfMeDIP-seq [133-135].

Intrigued by the low-input improvements in cfMeDIP-seq, we optimized MBD-seq [139] to enable as little as 1 ng cfDNA input and termed this ultra-low input protocol cfMBD-seq. We first optimized the methylation capture reaction in a 100 ng DNA mixture (10 ng cfDNA + 90 ng filler DNA) by adjusting the amount of MethylCap protein and protein-binding Dynabeads. We then tested the effect of the methylation status of filler DNA in the methylation enrichment. We also compared the methylome profile generated by cfMBD-seq across different DNA input (1-100 ng) with the one generated by standard MBD-seq (1000 ng DNA input). Additionally, we investigated if using more filler DNA, applying double wash, or using elution buffer with different salt concentration can further improve the methylation enrichment. To verify if the methylation signals captured by cfMBD-seq are real, we compared cfMBD-seq with most existing methylation profiling assays including HM450K, WGBS, reduced representation bisulfite sequencing (RRBS), and cfMeDIP-seq. All the results we showed below demonstrate that this novel bisulfite-free ultra-low input technology is promising in non-invasive, highly sensitive, and cost-effective methylation profiling.

2.2 Material and methods

2.2.1 DNA extraction

Pooled human plasma (IPLAWBK3E50ML) was purchased from Innovative Research (Novi, MI, USA). Whole blood (K3 EDTA tube) was collected from donors in an FDA-approved collection center. Plasma was frozen immediately after isolation. After thawing, additional centrifugation on 3000 rpm for 10 mins was performed to ensure complete depletion of cell debris. cfDNA was extracted using the QIAamp Circulating Nucleic Acid Kit (Qiagen; Hilden, Germany) and quantified with a Qubit Fluorometer using the iQuant™ NGS-HS dsDNA Assay Kit (Genecopoeia; Rockville, MD, USA). The average cfDNA yields from 1 ml plasma was ~7.5 ng. Colorectal carcinoma cell line HCT116 was purchased from ATCC (CCL-247™) and cultured according to the recommended cell culture method. HCT116 DNA was extracted using the QIAamp DNA Blood Mini Kit (Qiagen) and quantified with Nanodrop (NanoDrop Technologies; Wilmington, Delaware, USA). gDNA was sheared to 160 bp using a Covaris ME220 Focused-Ultrasonicator to mimic the fragment size of cfDNA. HCT116 was chosen because of the availability of public DNA methylation data.

2.2.2 Filler DNA generation

To generate filler DNA, Enterobacteria phage λ DNA was polymerase chain reaction (PCR) amplified with the GoTaq Master Mix (Promega; Madison, WI, USA). Primer sequences are as follows: Forward primer 5’- CGATGGGTTAATTCGCTCGTTGTGG-3’, reverse primer 5’-GCACAACGGAAAGAGCACTG-3’. The 274 bp amplicons were treated with CpG methyltransferase (M.SssI; Thermo Fisher Scientific) to methylate amplicons. Methylated amplicons were purified by the DNA Clean & Concentrator-5 Kit (ZYMO Research; Irvine, CA, USA) and quantified by Qubit Fluorometer. CpG methylation-sensitive restriction enzyme

HpyCH4IV (New England BioLabs; Ipswitch, MA, USA) digestion followed by agarose gel electrophoresis was performed to ensure complete methylation of filler DNA.

2.2.3 Library preparation

DNA was subjected to end repair/A-tailing and adapter ligation using the KAPA Hyper Prep Kit (Kapa Biosystems; Wilmington, MA, USA) with the sequencing adapter from NEBNext Multiplex Oligos for Illumina (New England BioLabs; Ipswitch, MA, USA). The amount of adapter used in the reaction was adjusted according to an adapter:insert molar ratio of 200:1. Adapter ligated DNA were purified with 0.8x SPRI Beads (Beckman Coulter; Pasadena, CA, USA) and digested with the USER enzyme (New England BioLabs) followed by purification with the DNA Clean & Concentrator-5 Kit (ZYMO Research; Irvine, CA, USA). Adapter ligated DNA was first combined with methylated filler DNA to ensure that the total amount of input for methylation enrichment was 100 ng, which was further mixed with 0.2 ng of methylated and 0.2 ng of unmethylated A. thaliana DNA from the DNA Methylation control package (Diagenode, Seraing, Belgium).

2.2.4 cfMBD-seq

The DNA mixture was then subjected to methylation enrichment using the MethylCap Kit (Diagenode) following the manufacture's protocol with some modifications. The total volume brought up by Buffer B was reduced from 141.8 µl to 136 µl to minimize DNA waste. The amount of MethylCap protein and magnetic beads were decreased proportionally according to the recommended DNA to protein and beads ratio (0.2 µg protein and 3 µl beads per 100 ng DNA input). Single fraction elution with the High Elution Buffer was applied. The eluted fraction was purified by the DNA Clean & Concentrator-5 Kit. The purified DNA was divided into two parts, one for qPCR (PowerUp™ SYBR™ Green Master Mix, Thermo Fisher) quality control, another for library amplification. The recovery of spiked-in methylated and unmethylated control can be calculated based on the cycle threshold (Ct) value of the enriched sample and input control. The specificity can be calculated by (1 - [recovery of unmethylated control DNA over recovery of methylated control DNA]) × 100). The methylation-enriched DNA libraries were amplified as follows: 95 °C for 3 min, followed by 12 cycles of 98 °C for 20 s, 65 °C for 15 s and 72 °C for 30 s and a final extension of 72 °C for 1 min. During the amplification, unique index from primer was added to the sequencing adapter for each sample. The amplified libraries were purified using 1x SPRI Beads followed by a dual size selection (0.6x followed by 1.2x) to remove any adapter dimers. All final libraries were first quantified with both Qubit assay and qPCR-based quantification using the KAPA Library Quantification Kit (Kapa Biosystems). Then they were submitted to Moffitt Cancer Center Molecular Genomics Core for the D1000 ScreenTape Assay (Agilent; Santa Clara, CA, USA). Sequencing was performed on the NextSeq 550 platform (Illumina; San Diego, CA, USA) with high-output 75 bp single-end read, multiplexed as ~12-15 samples per run. The sequencing data are available on GEO under accession number GSE161331.

2.2.5 Data processing

After sequencing, pre-alignment quality control was performed for the raw sequenced reads using fastp (Version 0.20.1) [155] with the default settings. The sequenced reads were then aligned to the human genome (hg19) using Bowtie-2 (Version 2.4.2) [166] with the default settings. The generated sam files were converted to bam files, followed by sorting, indexing, duplicate read removal, and read count extractions on chr1 - chr22 using SAMtools (Version 1.11) [216] 'sort', 'index', 'markdup', and 'view' command lines, respectively. R (Version 4.0.3 or greater) package RaMWAS (Version 1.12.0) [175] was used for quality control of the overall mapping quality and calculation of average non-CpG/CpG ratio and coverage by CpG density. To ensure the comparability between different conditions, bam files from the same experimental condition were merged together, and 30 million sequenced reads were randomly extracted (https://github.com/ACSoupir/MiscProcessingScripts) from each condition for the plotting of coverage by CpG density plot. R package MEDIPS (Version 1.40.0) [168] was then applied for saturation analysis and calculation of correlations of genome wide short read counts profiles between samples based on counts per 1000 bp non-overlapping window. Normalized data were exported as wiggle files for visualization on the Integrative Genomics Viewer [177].

Genome-wide CpG annotations reference was obtained from R package annotatr (Version 1.16.0). BEDtools (Version 2.28.0) [217] 'coverage' command line was used to call the depth of features according to the CpG annotations reference. TPM (Transcripts Per Kilobase Million) normalization was performed before comparing the read counts of CpG annotations regions between different samples. Data from low-input MBD-seq and cfMeDIP-seq were reprocessed from raw data (fastq level) using the same workflow. R package minfi (Version 1.36.0) was used to call and annotate (hg19) methylation signal from Infinium HM450K data. Average beta-values of each CpG site among different samples was first calculated. Methylation status of CpG islands was then determined by the average beta-values of adjacent CpG sites within the same CpG island (<0.5 as unmethylated and ≥0.5 as methylated). Logistic regression model was built using normalized read counts from cfMBD-seq and methylation status (methylated as 1 and unmethylated as 0) from microarray. R package ROCR (Version 1.0-11) was used to generate receiver operating characteristic curve. All data and R images were imported into GraphPad Prism 8 for the preparation of figures. Detailed bioinformatics analysis pipeline was coded in git bash and is available in GitHub (https://github.com/LiangWangLab/cfMBD-seq).

2.3 Results

2.3.1 Characterization of cfMBD-seq technology

The standard protocol for MBD methylation enrichment requires a minimum input of 1000ng DNA. Since the yield of cfDNA is extremely low (2-10 ng per ml plasma), the current protocol is not suitable for cfDNA methylation analysis. To guarantee amplification of methylation-enriched cfDNA, we added sequencing adapters to cfDNA by end repair/A-tailing and ligation before the methylation enrichment. Since newly synthesized DNA are not methylated, library amplification was not performed until the methylation enrichment is done. To meet the high input requirement for methylation enrichment, we added exogenous Enterobacteria phage λ DNA (filler DNA) to the adapter-ligated cfDNA to increase the final DNA input to 100 ng. The filler DNA ensures a constant MethylCap protein/DNA ratio and helps maintain a similar methylation enrichment efficiency across different samples with different cfDNA yields, while minimizing non-specific binding and DNA loss. Since the filler DNA is not adapter ligated, thus it is not amplified and sequenced, it will not interfere with the analysis of sequencing data. Beside filler DNA, we also added methylated and unmethylated A. thaliana DNA as spike-in controls to verify the specificity of methylation enrichment. Unlike genome wide sequencing, cfMBD-seq captures only a fraction of the genome (methylated DNA) and thus allows adequate sequencing depth from less total reads. Therefore, it enables pooling of multiple uniquely indexed samples for a single sequencing run. This makes cfMBD-seq a cost-effective method for methylome-wide association analysis in a large-scale study (for more details, see Methods and **Figure 5a**).

2.3.2 Reduced MethyCap protein improves methylation enrichment

Based on the use of filler DNA, we performed extensive benchmarking to identify an optimal methylation enrichment condition. One of the key adaptions for this purpose is to determine an appropriate amount of MethylCap protein to bind the input DNA mixture. If the amount of protein is too high, non-specific binding will occur due to extra binding sites on the protein. If too low, a portion of methylated fragments will not be captured. Using a mixture of 10 ng adapter ligated cfDNA and 90 ng filler DNA as input, we tested across different ratios of MethylCap protein and magnetic beads to the input DNA. When we kept the same MethylCap protein/DNA ratio as recommended by the manufacturer's protocol, where 2 μg MethylCap protein is used for 1 μg DNA (2:1 ratio), the captured CpG islands reached up to 58.65% of all mapped reads (**Figure 6a**). Since methylation differences sometimes occur at a short distance away from the CpG islands [218], we also calculated the sum of captured reads from CpG islands/shores/shelves regions. Under the recommended 2:1 ratio, 94.56% of reads were mapped into the extended regions, while these regions only account for 6.72% of the entire genome (**Figure 5b, 6b**). We then plotted the genome-wide coverage (average number of fragments covering CpGs) against CpG density (number of CpGs per fragment). The curve showed that the number of sequence reads was relatively low in CpG-poor regions and ultra-dense regions, while peaks in regions with moderate CpG density. As the peak represents CpG-rich regions such as CpG islands, the higher coverage at the peak indicates the better methylation enrichment (**Figure 6c**). To better characterize these distributions, we termed the CpG density at the point of the highest coverage as "peak". We also used the term "noise" to illustrate the ratio of average non-CpG/CpG coverage. Consistently, the 2:1 ratio gives the highest peak and the lowest noise values (**Figure 6d**). Unlike the tremendous impact from MethylCap protein, the amounts of magnetic beads had less impact on the methylation enrichment. Given redundant beads may increase the risk of unspecific binding, we determined 0.2 μg protein and 3 μl beads as the best enrichment condition for 100ng DNA input.

2.3.3 Methylated filler DNA is needed to increase enrichment efficiency

In MBD methylation enrichment, the typical yields of methylated DNA are 3-20% of the input DNA mass. Since cfDNA only accounts for a small fraction (<10%) in the mixture of cfDNA and filler DNA, the tiny amount of methylated cfDNA may not be able to fill all binding sites on the MethylCap proteins. If the filler DNA is not methylated, the risk of unspecific binding may be increased. To test the potential impact of filler DNA methylation status on the enrichment efficiency, we used different methylation status of filler DNA for cfMBD-seq, including: 1. Treated with CpG methyltransferase; 2. The mixture of the treated and untreated (1:1 ratio); and 3. Untreated. When the filler DNA is methylated, we observed better enrichment of sequence reads in both the CpG islands and CpG islands/shores/shelves regions. The sequence reads percentage of these regions were decreased with the reduction of filler DNA methylation level (**Figure 7a,7b**). Specifically, the percentage of sequence reads on CpG islands was 58.65%, 40.05%, and 20.53% when methylation level of filler DNA was 100%, 50%, and 0%, respectively. The extended regions showed the same trend. The coverage by CpG density plot (**Figure 7c**) and peak/noise plot (**Figure 7d**) further showed the importance of fully methylated filler DNA. Since the methylated filler DNA can block the extra binding sites on the MethylCap protein, it is not difficult to explain why the specificity of the reaction was enhanced.

2.3.4 Pre-sequencing quality controls

As we empirically found that the library yields from different experiment conditions are different, we hypothesized that a non-specific methylation capture would generate a higher library yield. To test this, we examined final library concentration and the quality of methylation enrichment. We found that the optimal condition tended to have a lower library yield while the suboptimal conditions generated more final library DNA when under the same PCR amplification cycles

(**Figure 8a, 8b**). This can be explained by the high specificity of the optimal methylation enrichment that only captures methylated DNA. Beside library concentration, real-time PCR (qPCR) often provides a more accurate pre-sequencing quality control. Since cfDNA is highly fragmented, the use of large amplicon is not recommended. Thus, the methylated control TSH2B (170 bp) provided in the kit is not an optimal spike-in control. On the other hand, qPCR has a limited sensitivity to detect the provided unmethylated control GAPDH in a successful enrichment due to low input. Therefore, instead of the TSH2B and GAPDH control pair provided in the kit, we used methylated and unmethylated A. thaliana DNA as spike-in controls to estimate the enrichment efficiency. We observed a significant enrichment of methylated DNAs when compared the spike-in controls before and after the capture reaction. Under the optimal condition, the specificity of capturing methylated control DNA was ⩾99%. Additionally, the recovery rate of the methylated and unmethylated control should be ~50%-90% and <1%, respectively. (**Figure 8c, 8d**).

2.3.5 1ng input achieved high quality results like 1000ng input DNA

To investigate if the methylome from low-input cfMBD-seq is equivalent to the one from standard MBD-seq (>1000 ng input), we sheared colorectal cancer cell line HCT116 DNA into small fragments with a peak of ~167 bp to mimic cfDNA and tested the efficiency of methylation enrichment across different DNA input (1, 10, 100, and 1000 ng). For 1 ng and 10 ng HCT116 DNA, we used methylated filler DNA to increase the final DNA input to 100 ng and followed the cfMBD-seq protocol. For 100 ng HCT116 DNA, we directly applied the cfMBD-seq protocol without filler DNA. For 1000 ng HCT116 DNA, standard MBD-seq protocol was used. Saturation analysis of sequencing result showed a high saturation correlation across different DNA input (**Figure 9a-d**). Specifically, the saturation correlation is 0.91 when the DNA input is only 1 ng, indicating that 1 ng DNA is sufficient to generate a saturated and reproducible coverage profile of the reference genome. The saturation correlations of 5 ng cfDNA input are consistent with the 10 ng genomic DNA (gDNA) input (**Figure 9e-f**). Additionally, the results showed robust genome-wide inter-replicate Pearson correlation (**Figure 10a**). Together, these results suggest that cfMBD-seq can generate a high quality methylome that is equivalent to standard MBD-seq while allowing ultra-low DNA input.

As the 1000 ng input has the highest genome-wide inter-replicate correlation, we further investigated if an increased amount of filler DNA can enhance the performance of the reaction. We thus increased the DNA input by adding more filler DNA, with the quantity of cfDNA unchanged (in total 100, 500, 1000 ng). However, we didn't observe an improved methylation enrichment even when the amounts of MethylCap protein and beads were adjusted accordingly. In fact, the higher amount of filler DNA reduced the performance of CpG-islands-centered methylation enrichment (**Figure 10b-c**) and increased background noise (**Figure 10d**). These results can be explained by the increased amount of methylated filler DNA overshadowed the trace amount of methylated cfDNA. Thus, we decided 100 ng as an optimal DNA input, due to the robust recovery of CpG-islands-centered regions with low noise.

2.3.6 Additional wash and elution did not improve enrichment

Given the confirmed MethylCap protein to DNA ratio and amount of methylated filler DNA, we evaluated other experimental conditions to see if the methylation enrichment performance can be further improved. First, we examined the effect of a more stringent wash condition on non-specific binding. However, double wash did not significantly increase the percentage of CpG island reads when compared to the standard wash. The additional wash also did not decrease the percentage of sequence reads on the open sea regions, where non-specific bindings are most likely to occur

(**Figure 11a**). Likewise, there was no significant difference on the peak and noise between the standard wash and double wash (**Figure 11b**). Since the additional wash can take much more time, we will not consider it as an optimization.

Second, we examined the effect of the elution buffer salt concentration on methylation enrichment. We performed single fraction elution using the three different elution buffers (High, Medium, Low) provided in the MethylCap kit. Theoretically, an increased salt concentration may preferentially enrich regions with higher CpG density [219]. However, we did not observe a notable shift on the coverage by density plot, nor sequence reads percentage difference on each CpG annotations (**Figure 12a-c**). For example, the signals at the CpG island *MGAT3* showed no difference among the three elution buffers (**Figure 12d**). The finding that MethylCap protein (MeCP2) is not sensitive to the salt concentration of elution buffer has been reported previously [220,221]. We also investigated multiple fractions elution, that is, sequential elution with low, medium, and high salt elution buffer from one capture reaction. The coverage by density plot illustrated robust methylation enrichment in both the first fraction (low salt elution buffer) and the pool of three fractions (**Figure 12e**). But the second fraction (medium salt), the third fraction (high salt), and the pool of the second and third fractions had very low coverage (**Figure 12e**). These results demonstrated the importance of the first fraction of elution, no matter the salt concentration of elution buffer, due to the intrinsic limitation of ultra-low DNA input. In summary, our results suggest an optimal condition for low input methylation enrichment includes 0.2 μg MethylCap protein and 3 μl beads for 100 ng DNA mixture (cfDNA + methylated filler DNA), standard wash, and single fraction elution.

2.3.7 Comparison of cfMBD-seq with other technologies

To evaluate the methylation capture accuracy of cfMBD-seq, we calculated its sensitivity (proportion of methylated CpG islands detected) and specificity (1 - proportion of non-methylated CpG islands detected). We used Infinium HM450K data (Gene Expression Omnibus (GEO): GSE55491, peripheral blood mononuclear cell (PBMC) from N=5 healthy controls) as a standard to determine whether a CpG island was methylated or non-methylated. It is known that the methylation level between neighboring CpG sites is positively correlated. Therefore, to obtain a comparable measurement between cfMBD-seq and methylation array, we averaged the beta-values of adjacent CpG sites within each CpG island and defined the methylation status of that CpG island. We then built a logistic regression model for all CpG islands on the microarray using normalized read counts from cfMBD-seq and methylation status from microarray (AUC=0.995, **Figure 13a**). At the cutoff of 13.25, derived from the intersection of the specificity and sensitivity curves translated to normalized read counts, the sensitivity of cfMBD-seq is 0.94 and the specificity is 0.98. Namely, at this threshold, cfMBD-seq detected 94% of the methylated CpG islands that were reliably detected by Infinium methylation array, while correctly classifying 98% of non-methylated sites.

To determine the performance of cfMBD-seq over existing methylation enrichment assays, we compared cfMBD-seq with a previously published low input MBD-seq protocol (N=4 from GEO: GSM2593327-GSM2593330) that did not use filler DNA [141]. Different from cfMBD-seq, this low input MBD-seq protocol has a very low recovery rate of the CpG island regions (median 19.95% [(Q1) 19.25%-(Q3) 20.11%]) and a relatively high recovery rate of the open sea regions (14.30% [14.24%-14.49%]) (**Figure 13b-c**). Worst of all, the overall coverage is low, which makes it difficult to discriminate methylated fragments from non-specific fragments (**Figure 13d**). We next compared cfMBD-seq with cfMeDIP-seq (N=24 cancer-free individuals from published dataset) which showed adequate performance on capturing tumor-specific methylation in cfDNA [133-135]. According to the summary QC from the RaMWAS package, we observed a higher percentage of reads that passed the filter in cfMBD-seq (83.15% [82.93%-83.68%]) than in cfMeDIP-seq (74.90% [74.53%-75.45%]) and a lower duplicate rate (3.45% [3.40%-3.90%] vs. 12.00% [9.00%-19.23%]). Taken together, cfMBD-seq generated higher quality and more informative sequencing data than cfMeDIP-seq (79.60% [79.15%-80.43%] vs. 62.65% [55.60%-66.65%]) (**Table 1**). For the CpG annotation-based results, cfMBD-seq showed a significantly higher recovery rate at CpG islands (60.13% [58.78%-60.81%] vs. 38.16% [37.21%-41.28%], **Figure 13b**) and a slightly higher recovery rate at combined CpG islands/shores/shelves (94.81% [94.61%-94.98%] vs. 90.90% [90.91%-91.55%], **Figure 13c**), suggesting that cfMBD-seq preferentially enriches CpG islands, while cfMeDIP-seq has more signal on CpG shores and CpG shelfs. This finding is consistent with the coverage by CpG density plot, where cfMBD-seq peaks at higher CpG density than cfMeDIP-seq (29.98 [29.54-30.33] vs. 22.88 [22.37-23.50], **Figure 13d**). The comparison between cfMBD-seq, low input MBD-seq, and cfMeDIP-seq was summarized in **Table 1**. To better demonstrate the reproducibility of cfMBD-seq, we showed a snapshot of a genomic region with consecutive CpG islands (chr8:86,703,816-86,880,439). We observed peaks with high similarity among cfMBD-seq (1 to 100 ng input DNA), standard MBD-seq (1000 ng), cfMeDIP-seq (1 to 10 ng), and standard MeDIP-seq (100 ng) (**Figure 14**). We then compared the signal peaks among different methylation profiling technologies. We showed that cfMBD-seq also recapitulated methylation profiles from RRBS (1000 ng) and WGBS (2000 ng) (**Figure 15**). All these findings suggest that cfMBD-seq, allowing ultra-low DNA input as starting material, can reliably detect genome-wide DNA methylation signal with high accuracy. These features demonstrate that cfMBD-seq is a promising tool in the discovery of novel biomarkers for cancer detection and management.

2.4 Discussion

In this study, we further optimized the MBD-seq protocol to enable methylation enrichment from ultra-low DNA input. The most critical modification for this purpose is to decrease the amount of MethylCap protein proportionally according to the MethylCap protein/DNA ratio recommended by the manufacturer's protocol. The 2:1 MethylCap protein/DNA ratio ensures a high specificity of methylation capture (**Figure 6**). Unlike MethylCap protein, the amounts of magnetic beads had less impact on the methylation enrichment. By comparing across different methylation status of filler DNA, we showed that methylated filler DNA is also indispensable to ensure the specificity of methylation capture (**Figure 7**). Using as little as 1 ng DNA input, cfMBD-seq is able to generate a saturated and reproducible coverage profile of the reference genome (**Figure 9**). Also, the methylome from 1 ng DNA generated by cfMBD-seq is highly correlated to the methylome from 1000 ng DNA generated by standard MBD-seq (**Figure 10**). However, other attempts such as using more filler DNA, applying double wash, or using elution buffer with different salt concentration cannot improve the performance of methylation enrichment (**Figure 11,12**). To evaluate if the methylation signals captured by cfMBD-seq are real, we compare our sequencing data with the most commonly used HM450K assay by a logistic regression model. The results showed that cfMBD-seq detected 94% of the methylated CpG islands detected by HM450K, while correctly classifying 98% of non-methylated sites (AUC=0.995) (**Figure 13a**). cfMBD-seq also performs better than a previously published low input MBD-seq protocol in the methylation enrichment of CpG islands-centered regions [141]. Most importantly, cfMBD-seq outperforms cfMeDIP-seq in the enrichment of fragments with higher CpG density such as CpG islands (**Figure 13b-d**).

The differences between standard MBD-seq and standard MeDIP-seq have been described in a previous study: MeDIP commonly enriches methylated regions with a low CpG density, while MBD captures a broad range of CpG densities and identifies the greatest proportion of CpG islands [140]. It is known that CpG-rich fragments do not undergo complete denaturation into single stranded DNA which is required for an efficient MeDIP capture and may explain why MeDIP-seq is less sensitive toward fragments with high CpG density. In contrast, MBD capture does not require DNA denaturation because the MethylCap protein is sensitive toward double stranded DNA. Therefore, temperature control of DNA-protein mixture during MBD capture is less strict than MeDIP capture. In addition, MBD enrichment in cfMBD-seq can be finished within 5 hours (including 3 hours of incubation), while cfMeDIP enrichment needs overnight incubation. Thus, the reaction of MBD enrichment is less time-consuming. Overall, both cfMBD-seq and cfMeDIP-seq are reliable ultra-low input methylation profiling assays. cfMBD-seq is a method of choice for interrogating regulation of gene expression (methylation changes in CpG islands), while cfMeDIP-seq would be preferable in investigating transcriptional regulation of non-coding RNAs (methylation changes in gene bodies and CpG shores).

There are a few caveats to ensure successful cfMBD-seq. First, the quality of the MethylCap protein is very important. We notice that the use of MethylCap protein that has experienced multiple freeze–thaw cycles can negatively impact the data quality. Because the MethylCap protein is used with 10-fold dilution before adding to the reaction, it can be used for much more reactions than the standard MBD capture. Therefore, we recommend splitting the MethylCap protein into multiple aliquots to minimize the freeze-thaw cycles and using fresh diluted protein for each batch. Second, the success of the methylation enrichment reaction must be validated by qPCR to detect recovery of spiked-in control. The specificity of the reaction should be ⩾99% before proceeding to the next step. Third, accurate library quantification is critical. As methylated filler DNA is used in the methylation enrichment, fluorometer-based library quantification methods may inevitably count filler DNA. Therefore, qPCR-based library quantification is recommended because it will only quantify the amounts of amplifiable DNA (adapter ligated cfDNA). Lastly, adequate sequencing depth is crucial for high quality data. Based on the saturation analysis, at least 30 million mapped reads are required to generate a saturated and reproducible coverage profile. The cost of the entire cfMBD-seq workflow, starting from cfDNA extraction through the generation of sequencing data (single-end and pooling 12-15 indexed libraries using the Illumina NextSeq 550 platform), is less than $300 per sample. This cost-effective feature allows large-scale methylome-wide association analysis that is crucial for the establishment of a prediction model with high accuracy.

It is worth mentioning that the current study also has some limitations. First, it is well known that the methylation status is different between individuals. The differences observed among cfMBD-seq, low-input MBD-seq, and cfMeDIP-seq could be partly attributed to the difference in different plasma samples that were used. Thus, further validation is required. Second, the main application of cfMBD-seq is to identify cancer biomarkers in cfDNA. However, current study was limited to technology development and optimization. Further study in patient's plasma samples is needed to test the feasibility of cfMBD-seq in clinical settings, in particular to elaborate how well this technology can differentiate the tumor-derived cfDNA methylation signals from the high background cfDNA from normal blood cells.

Overall, our study demonstrates the potential benefits of using cfMBD-seq to profile the methylome of cfDNA with ultra-low DNA input. Current results provide justification for further validation using case and control plasma samples from different malignancies to perform differential methylation analyses. Since enrichment-based methods are analyzed by comparing the relative abundance of sequenced fragments, cfMBD-seq shares similar data analysis workflows with cfMeDIP-seq for the identification of DMRs and other downstream machine learning analyses. Another potential for cfMBD-seq is the use in other methylome-wide investigations that are limited by DNA yield. We confidently believe that cfMBD-seq, being non-invasive and cost-effective, has a great potential in identifying biomarkers for cancer detection and management.

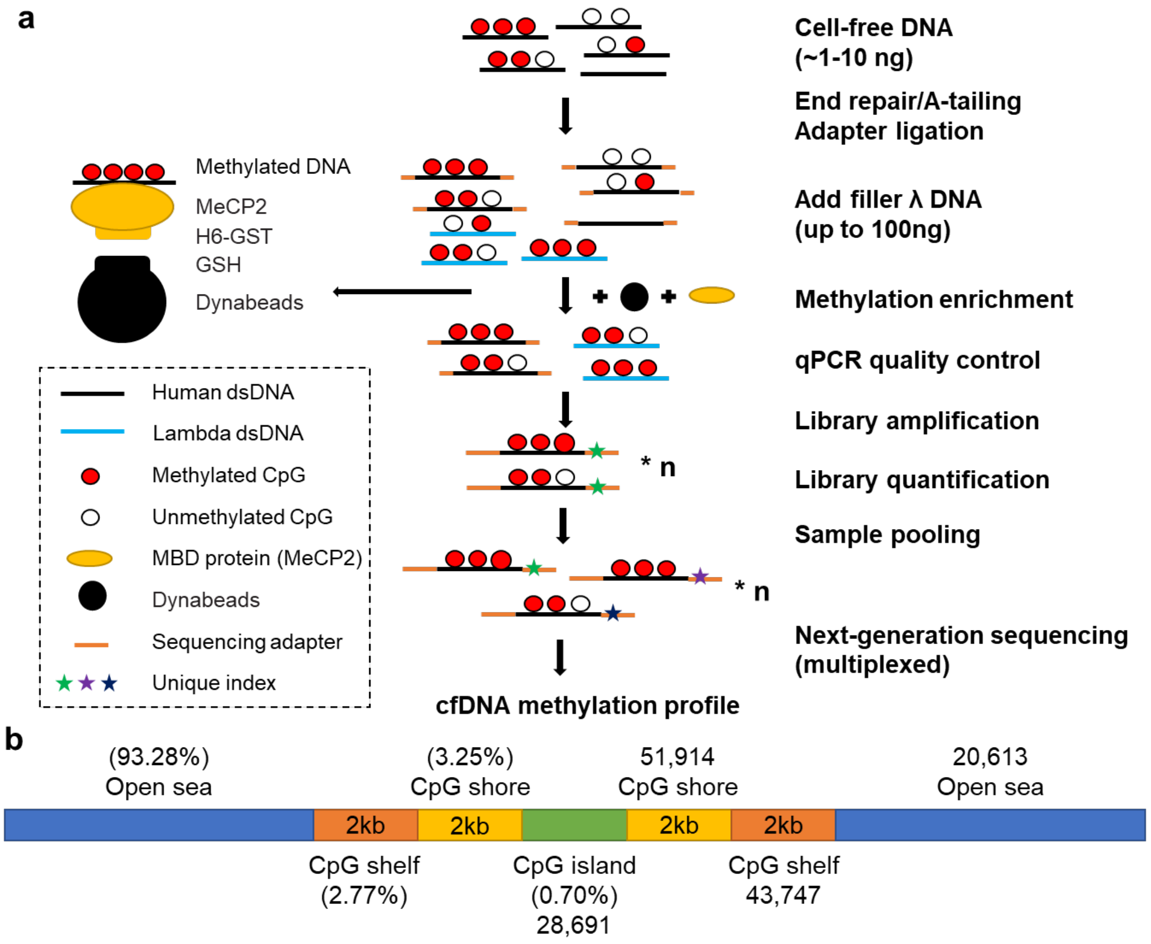


Figure 5. Schematic diagram of cfMBD–seq and CpG annotations

a) Schematic workflow of cfMBD–seq protocol. From cfDNA extraction to generation of methylation profile. b) Schematic diagram of CpG annotations. Numbers on the left (in brackets) represent the percentage of the CpG features in the human genome. For example, CpG islands account for only 0.7% of the human genome. Numbers on the right represent total number of features. For example, there are 28,691 CpG islands in the hg19 reference genome.

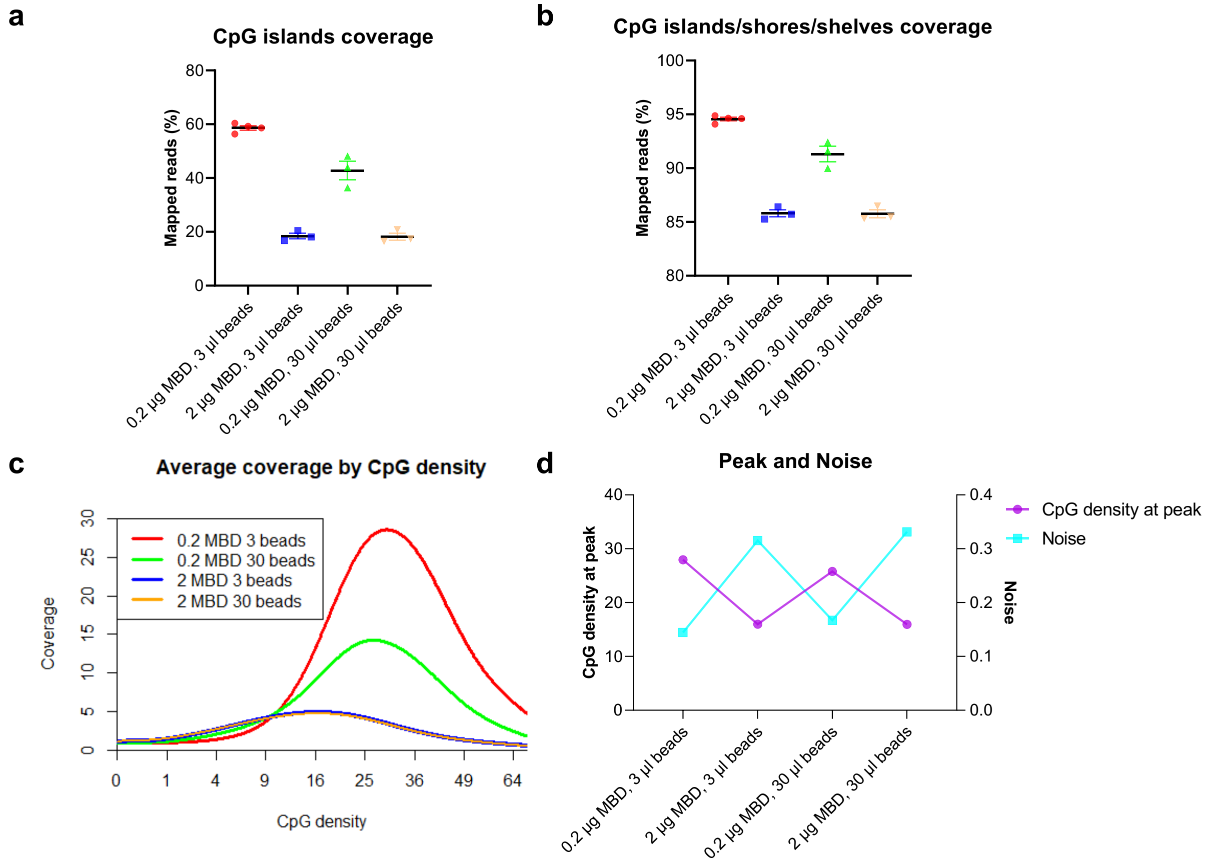


Figure 6. Reduced MethyCap protein improves low-input methylation enrichment

a) & b) Percentage of sequence reads mapped on CpG islands and CpG islands/shores/shelves across different amount of MethyCap protein and magnetic beads. (N=4 for the first condition, N=3 for other conditions. Mean with the standard error of the mean (SEM).) **c)** Coverage by CpG density plot across different amount of MethyCap protein and magnetic beads. Coverage is defined as average number of fragments covering CpGs. CpG density is number of CpGs per fragment. **d)** CpG density at peak and noise under different MethyCap protein and magnetic beads. CpG density at peak is CpG density at the point of highest coverage on the 'coverage by CpG density plot' (left y-axis). Noise is the ratio of average non-CpG coverage to average CpG coverage (right y-axis).

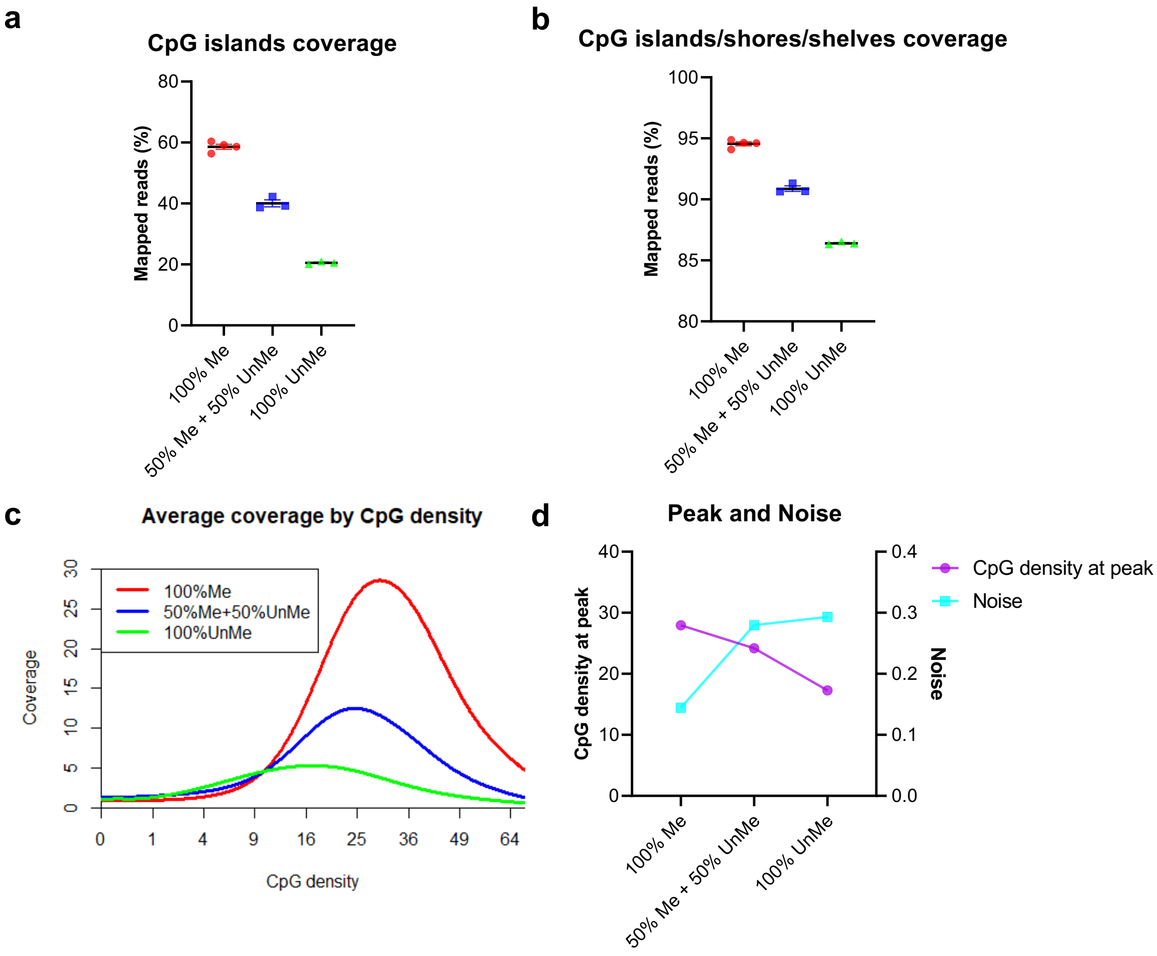


Figure 7. Methylated filler DNA is needed for low input methylation enrichment

a) & b) Percentage of sequence reads mapped on CpG islands and CpG islands/shores/shelves across different methylation status of filler DNA. (N=4 for the first condition, N=3 for other conditions. Mean with SEM.) c) Coverage by CpG density plot across different methylation status of filler DNA. d) The CpG density at peak (left y-axis) and noise (right y-axis) of different methylation status of filler DNA.

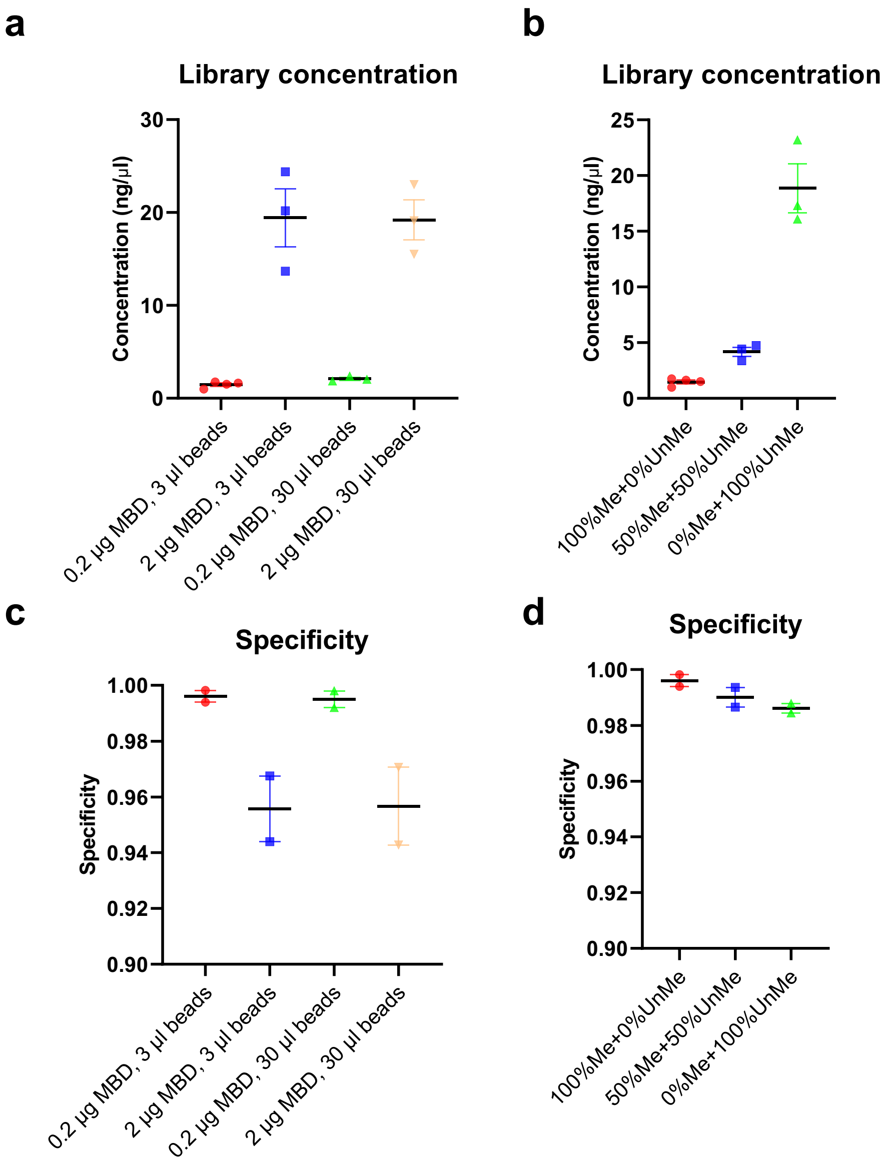


Figure 8. Important pre-sequencing quality controls

a) & b) Library concentration (ng/µl) measured by Qubit assay across different conditions. c) & d) Specificity of methylation enrichment measured by qPCR, using methylated and unmethylated spiked-in A. thaliana DNA control.

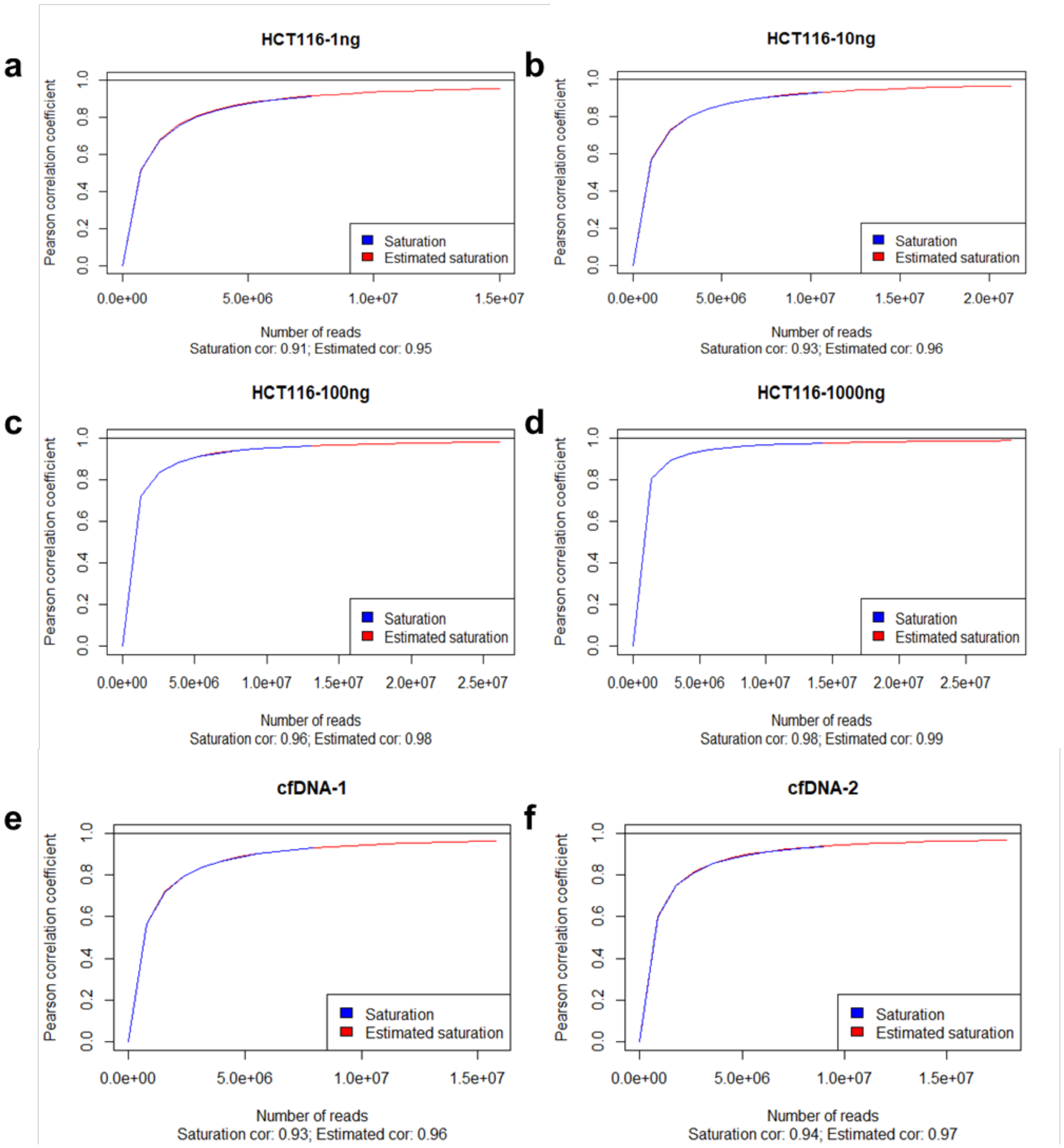
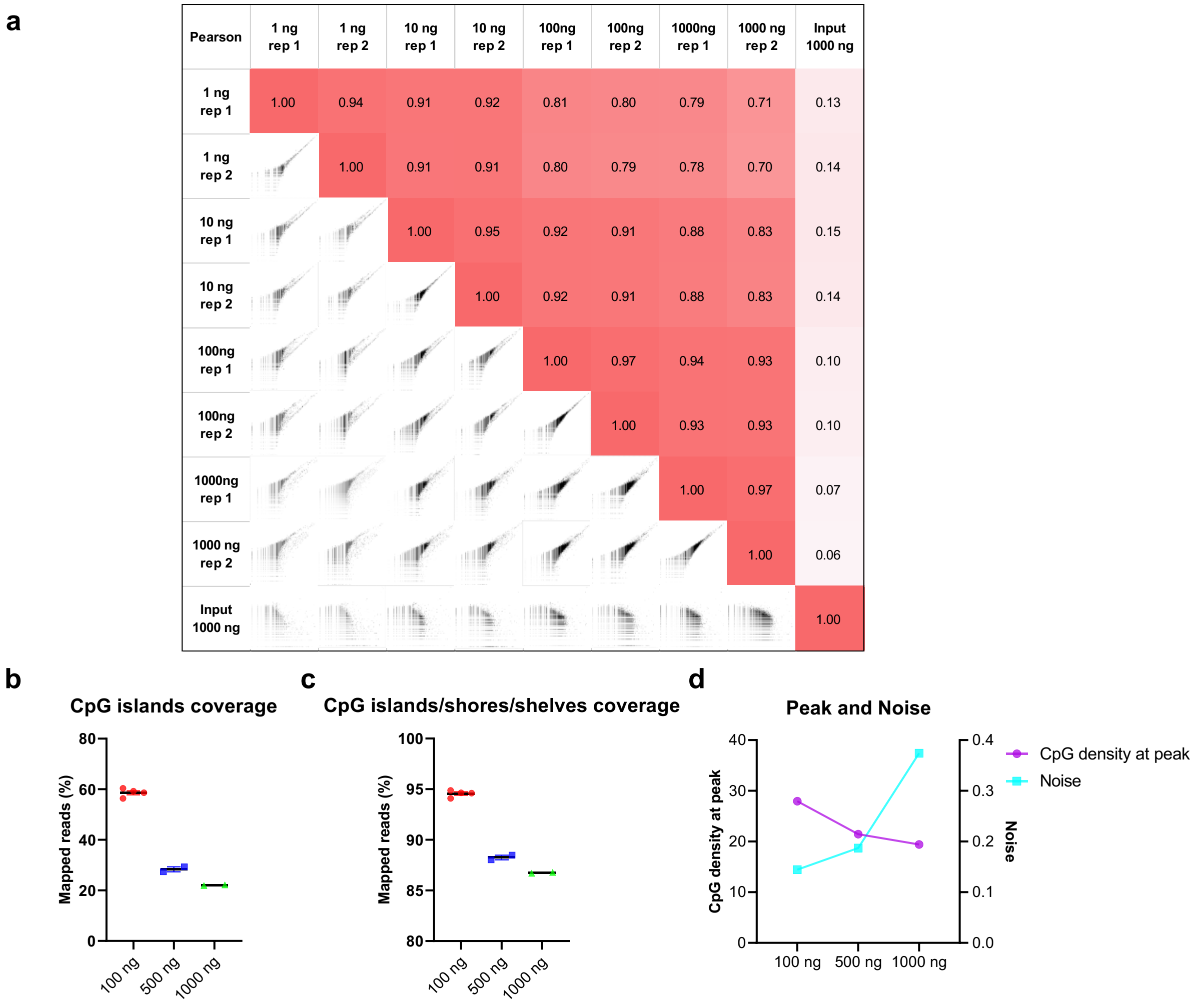


Figure 9. Saturation analysis across different DNA input

Saturation analysis from the MEDIPS R package for the sequencing result of cfMBD-seq using different HCT116 DNA input (1, 10, 100, and 1000 ng) (a-d) and 3 ng cfDNA input (e-f). The saturation analysis determines if the given set of mapped reads is sufficient to generate a saturated and reproducible coverage profile of the reference genome.



Pearson	1 ng rep 1	1 ng rep 2	10 ng rep 1	10 ng rep 2	100ng rep 1	100ng rep 2	1000ng rep 1	1000 ng rep 2	Input 1000 ng
1 ng rep 1	1.00	0.94	0.91	0.92	0.81	0.80	0.79	0.71	0.13
1 ng rep 2		1.00	0.91	0.91	0.80	0.79	0.78	0.70	0.14
10 ng rep 1			1.00	0.95	0.92	0.91	0.88	0.83	0.15
10 ng rep 2				1.00	0.92	0.91	0.88	0.83	0.14
100ng rep 1					1.00	0.97	0.94	0.93	0.10
100ng rep 2						1.00	0.93	0.93	0.10
1000ng rep 1							1.00	0.97	0.07
1000 ng rep 2								1.00	0.06
Input 1000 ng									1.00

Figure 10. Different amount of input DNA in cfMBD-seq

a) Genome-wide Pearson correlations of normalized read counts between cfMBD-seq signal for 1-1000 ng of input HCT116 DNA (2 technical replicates per concentration). The input control is from an input library of a ChIP-seq study (ENCODE: ENCFF280GWX). Log transformed counts were used in the scatter plots. b) & c) Percentage of sequence reads mapped on CpG islands and CpG islands/shores/shelves across different mixture of cfDNA and filler DNA. (N=4 for the first condition, N=2 for other conditions. Mean with SEM.) d) CpG density at peak (left y-axis) and noise (right y-axis) of different mixture of cfDNA and filler DNA.

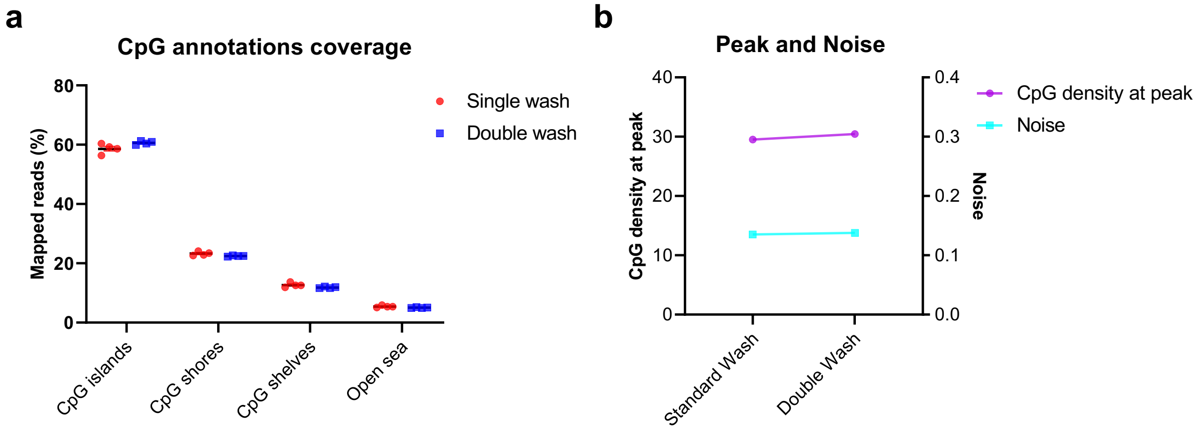


Figure 11. Additional wash does not improve methylation enrichment

a) Percentage of sequence reads mapped on different CpG annotations across different wash conditions. (N=4 for each condition. Mean with SEM.) b) CpG density at peak (left y-axis) and noise (right y-axis) of different wash conditions.

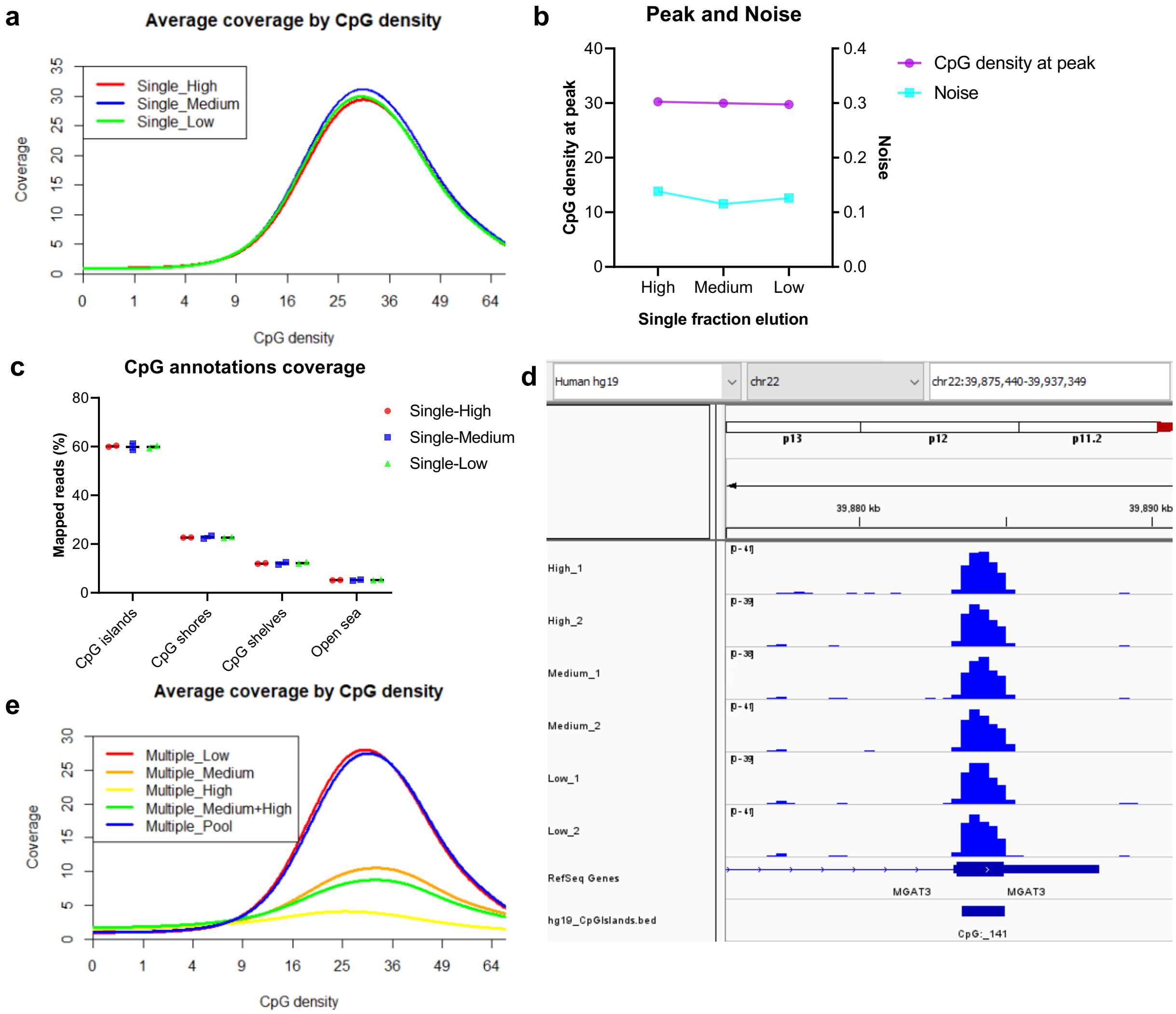


Figure 12. Effect of elution buffer in cfMBD capture

a) Coverage by CpG density plot across elution buffers with different salt concentration. b) CpG density at peak (left y-axis) and noise (right y-axis) of different elution buffers. c) Percentage of sequence reads mapped on different CpG annotations across different wash conditions. (Mean with SEM.) d) Genome Browser snapshot of cfMBD-seq signal at the CpG island of MGAT3, which is used as an example in the manual of MethylCap kit. Data were processed by MEDIPS package for RPKMs normalization and were exported as wiggle files for visualization. e) Coverage by CpG density plot across multiple fractions of elution conditions.

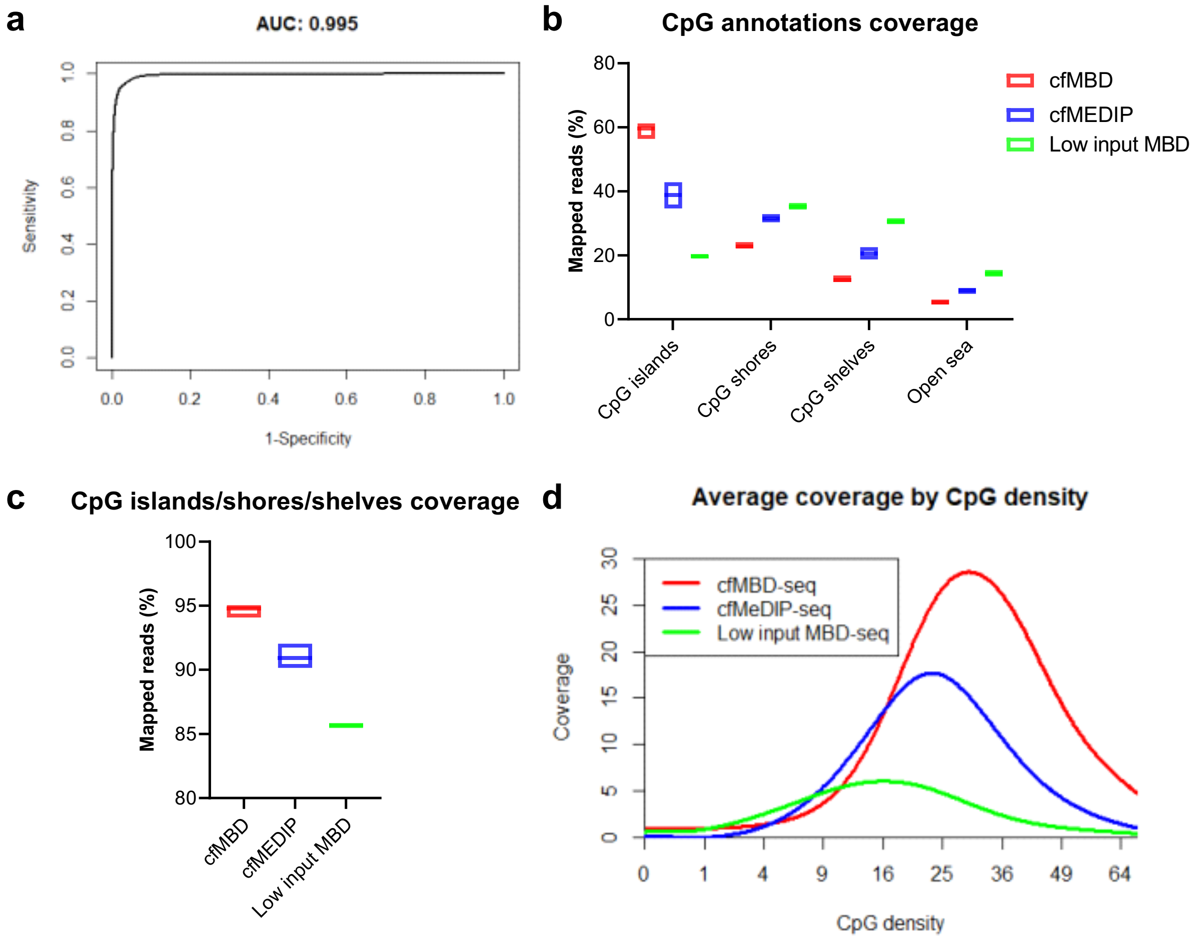


Figure 13. Comparison of cfMBD-seq with low input MBD-seq and cfMeDIP-seq

a) Receiver operating characteristic curve and corresponding area under the ROC curve for methylation status of CpG islands from Infinium HM450K data predicted by cfMBD-seq normalized read counts. b) & c) Percentage of sequence reads mapped on different CpG annotations (b) and CpG islands/shores/shelves (c) of cfMBD-seq (N=8), cfMeDIP-seq (N=24), and low input MBD-seq (N=4). (Mean with SEM.) d) Coverage by CpG density plot of cfMBD-seq, cfMeDIP-seq, and low input MBD-seq.

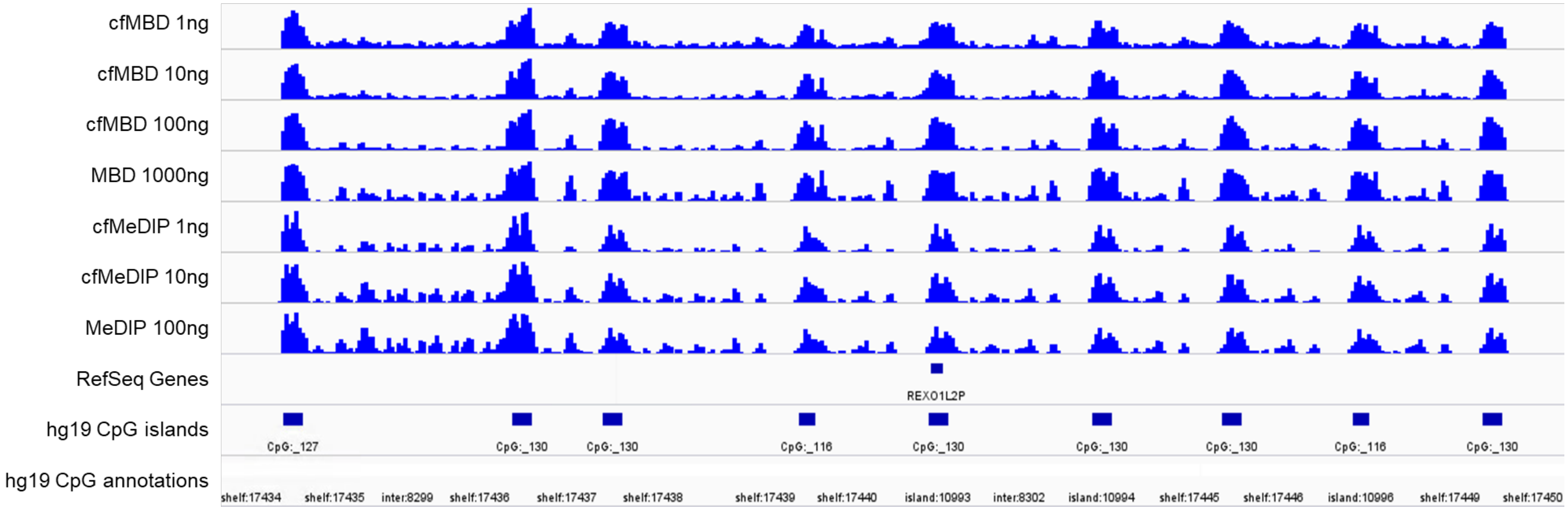


Figure 14. cfMBD-seq shares similar methylation profile with cfMeDIP-seq

Genome Browser snapshot of different input of HCT116 DNA signal by cfMBD-seq and cfMeDIP-seq across a region with consecutive CpG islands (chr8:86,703,816-86,880,439). Data were processed by MEDIPS package for RPKMs normalization and were exported as wiggle files for visualization.

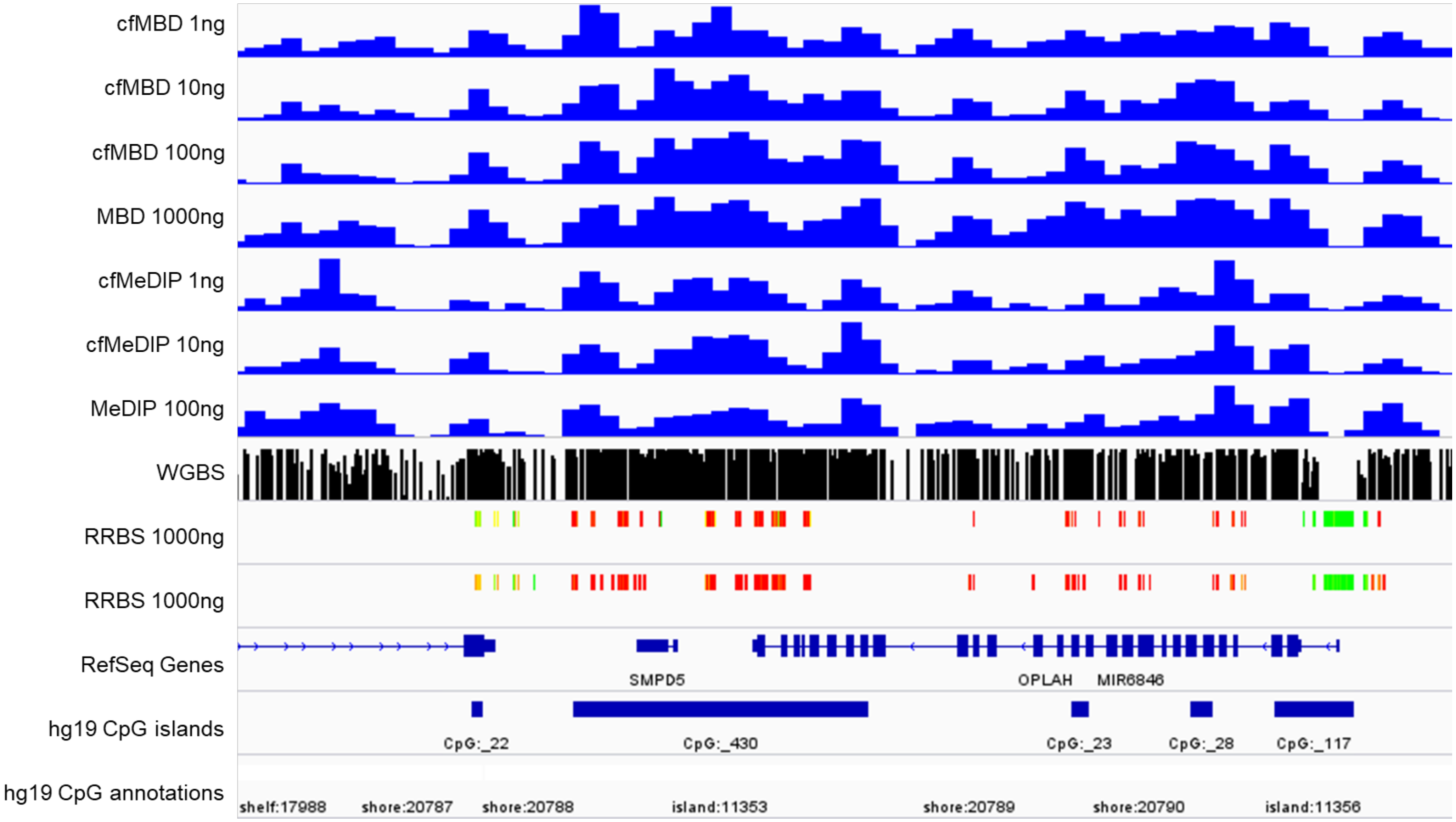


Figure 15. cfMBD-seq recapitulates methylation profiles from other technologies

Genome Browser snapshot of HCT116 cfMBD-seq signal across chr8:145,095,942-145,116,942, at different starting DNA input (1 to 100 ng), compared with cfMeDIP-seq (Gene Expression Omnibus (GEO): GSE79838), RRBS (ENCODE: ENCSR000DFS), and WGBS (GEO: GSM1465024) data. For cfMBD-seq and cfMeDIP-seq, the y axis indicates RPKMs normalized reads; for RRBS, red and green blocks represent hypermethylated and hypomethylated CpGs, respectively. For WGBS track, peak heights indicate methylation level.

Table 2. Comparison among cfMBD-seq, Low input MBD-seq, and cfMeDIP-seq

	cfMBD-seq (N=8)	Low input MBD-seq (N=4)	cfMeDIP-seq (N=24)
Experiment			
Filler DNA	Methylated DNA only	No filler	Mixture of methylated and unmethylated DNA
DNA Denaturation	Not required	Not required	Required
Capture protein	MeCP2	MBD2	Anti-5mc
Capture time	5 hours (including 3 hours incubation)	5 hours (including 3 hours incubation)	23 hours (including 17 hours overnight incubation)
Quality Control			
Reads passed filter	83.15% [82.93%-83.68%]	85.40% [85.03%-85.70%]	74.90% [74.53%-75.45%]
Duplicate rate	3.45% [3.40%-3.90%]	2.65% [2.60%-2.78%]	12.00% [9.00%-19.23%]
Used reads	79.60% [79.15%-80.43%]	82.75% [82.25%-83.10%]	62.65% [55.60%-66.65%]
Methylation Enrichment			
Reads on CpG islands	60.13% [58.78%-60.81%]	19.95% [19.25%-20.11%]	38.16% [37.21%-41.28%]
Reads on CpG islands/shores/shelves	94.81% [94.61%-94.98%]	85.70% [85.51%-85.76%]	90.90% [90.91%-91.55%]
CpG density at peak	29.98 [29.54-30.33]	15.76 [15.41-15.88]	22.88 [22.37-23.50]

Median along with [first quartile (Q1) - third quartile (Q3)] are shown.

Chapter 3: Cancer detection and classification by hypermethylated CpG islands

This section was previously published by Cancers, a peer-reviewed, open access journal of oncology, published semimonthly online by MDPI.

[222] Huang, J.; Soupir, A.C.; Schlick, B.D.; Teng, M.; Sahin, I.H.; Permuth, J.B.; Siegel, E.M.; Manley, B.J.; Pellini, B.; Wang, L. Cancer Detection and Classification by CpG Island Hypermethylation Signatures in Plasma Cell-Free DNA. Cancers (Basel) 2021, 13, doi:10.3390/cancers13225611.

3.1 Introduction

Lung and colorectal cancer are among the most common causes of cancer-related deaths in the US, while pancreatic cancer is the deadliest form of solid malignancy with an alarming 10% five-year survival rate [223]. Detection of cancer at an early stage before metastasis can significantly reduce the mortality of these malignances. Methylation in cfDNA is a promising biomarker for the early detection of cancer. CpG islands-specific hypermethylation is a key characteristic of the cancer methylome [45]. Hypermethylation of CpG island can affect the cell cycle, DNA repair, metabolism, cell-to-cell interaction, apoptosis, and angiogenesis, all of which are involved in tumorigenesis and cancer progression [224]. CpG island hypermethylation has been described in almost every tumor type [45]. One of the most well-studied DNA methylation signatures is the methylation of *SEPT9* promoter, which is an FDA-approved biomarker for colorectal cancer detection [225]. A blood-based test for methylated *SEPT9* (Epi proColon) has been applied to plasma cfDNA in patients undergoing colorectal cancer screening, however this test has a relatively low sensitivity for the detection of early-stage colorectal cancer [226]. Additionally, Epi proLung is another plasma-based DNA methylation test that detects methylated *PTGER4* and

SHOX2 promoters for the detection of lung cancer. Most recently, GRAIL lunched a methylation-based blood test called Galleri that could detect a range of cancers. The trail results from in a small number of studies showed that Galleri was able to detect over 50 types of cancer even at an early stage. Galleri has so far been promising, but it needs to be tested further in larger trials. Overall, CpG island hypermethylation has demonstrated its great versatility and potential for the detection and management of cancer [227].

Enrichment-based methylation profiling methods such as methyl-CpG-binding domain sequencing (MBD-seq) and methylated DNA immunoprecipitation sequencing (MeDIP-seq) have shown similar sensitivity and specificity for the detection of differentially methylated regions (DMRs) when compared to bisulfite conversion-based methods [131]. Nonetheless, such technologies are restricted to tumor tissue application due to the need of high amounts of DNA input. To address this issue, Shen et al. optimized the MeDIP-seq protocol to allow methylome analysis of small quantities of cfDNA, termed cfMeDIP-seq [134,135]. cfMeDIP-seq has shown high accuracy in the classification of a wide variety of cancer types [134] and characterization of renal cell carcinoma patients across all stages [136,137]. To expand the use of enrichment-based methods in cfDNA, we optimized the MBD-seq protocol for low input cfDNA methylation profiling, termed cfMBD-seq [101]. We previously showed that cfMBD-seq provides higher sequencing data quality with more sequenced reads passing filter and a lower duplicate rate than cfMeDIP-seq. cfMBD-seq also outperforms cfMeDIP-seq in the enrichment of high CpG density regions (i.e., CpG islands) [101]. However, the clinical feasibility of cfMBD-seq is unknown. Based on our previous findings, we hypothesized that cfMBD-seq can identify hypermethylated CpG islands as biomarkers for cancer detection and classification. In this study, we applied cfMBD-seq to the plasma samples of patients with advanced lung, colorectal, and pancreatic cancer and cancer-free individuals to determine whether cfMBD-seq can reliably identify differentially methylated regions (DMRs) between cases and controls. We also investigated whether these DMRs enable accurate discrimination between different cancer types (**Figure 16**).

3.2 Material and methods

3.2.1 Sample acquisition and clinical cohort

The study subjects were recruited at Moffitt Cancer Center following Total Cancer Care protocol (https://moffitt.org/research-science/total-cancer-care/). A total of 53 subjects including colorectal (N=13), lung (N=12), pancreatic (N=12) cancer patients, and non-cancer controls (N=16) were used in this study (Clinical demographic characteristics in **Table 3**). All cancer patients had metastatic disease at the time of sample collection. Most cancer patients had adenocarcinoma histology: 11 of 13 were colorectal adenocarcinoma; 9 of 12 were lung adenocarcinoma; and 10 of 12 were pancreatic adenocarcinoma. Subjects in the non-cancer cohort were specifically negative for any form of cancer. Samples were randomized and blinded during cfDNA extraction, library preparation, and sequencing. Samples were unblinded during data analysis. All patients provided written informed consent. The study was approved by Institutional Review Boards (IRB# 00000971) of H. Lee Moffitt Cancer Center & Research Institute (MCC 20563).

3.2.2 Plasma sample collection

Moffitt Cancer Center Total Cancer Care followed standard operating procedure for blood sampling: Whole blood specimens were obtained by routine venous phlebotomy and collected in Purple top EDTA blood tubes. Plasma was isolated from whole blood at the time of subject enrollment. Centrifugation of whole blood was performed at 1300 x g for 10min at room temperature. Plasma layer was transferred into 1.5 ml cryovials and stored as three 1mL aliquots. Plasma samples were frozen immediately at -80 °C after isolation.

3.2.3 cfDNA extraction

Plasma samples were thawed and centrifuged at 3,000g for 15 mins to ensure complete depletion of cell debris. cfDNA was extracted using the QIAamp Circulating Nucleic Acid Kit (Qiagen; Hilden, Germany) following the manufacturer's protocol, except for the addition of carrier RNA in Buffer AVE. All cfDNA eluates were quantified by a Qubit Fluorometer using the iQuant™ NGS-HS dsDNA Assay Kit (Genecopoeia; Rockville, MD, USA) and then submitted to Moffitt Cancer Center Molecular Genomics Core for D1000 ScreenTape Assay (Agilent; Santa Clara, CA, USA) to ensure the absence of high molecular weight DNA contamination from white blood cell lysis.

3.2.4 Filler DNA generation

To generate filler DNA, Enterobacteria phage λ DNA was polymerase chain reaction (PCR) amplified with the GoTaq Master Mix (Promega; Madison, WI, USA). Primers sequences are as follows: Forward primer 5'- CGATGGGTTAATTCGCTCGTTGTGG-3', reverse primer 5'-GCACAACGGAAAGAGCACTG-3'. The 274 bp amplicons were treated with CpG methyltransferase (M.SssI; Thermo Fisher Scientific) to methylate amplicons. Methylated amplicons were purified by the DNA Clean & Concentrator-5 Kit (ZYMO Research; Irvine, CA, USA) and quantified by Qubit Fluorometer. CpG methylation-sensitive restriction enzyme HpyCH4IV (New England BioLabs; Ipswitch, MA, USA) digestion followed by agarose gel electrophoresis was performed to ensure complete methylation of filler DNA.

3.2.5 Library preparation

cfDNA was subjected to end repair/A-tailing and adapter ligation using the KAPA Hyper Prep Kit (Kapa Biosystems; Wilmington, MA, USA) with the sequencing adapter from NEBNext Multiplex Oligos for Illumina (New England BioLabs). The amount of adapter was adjusted to an adapter:

insert molar ratio of 200:1. Adapter-ligated DNA were purified with 0.8 x SPRI Beads (Beckman Coulter; Pasadena, CA, USA) and digested with USER enzyme (New England BioLabs) followed by purification with the DNA Clean & Concentrator-5 Kit. Adapter ligated DNA was first combined with methylated filler DNA to ensure that the total amount of input for methylation enrichment was 100 ng, which was further mixed with 0.2 ng of methylated and 0.2 ng of unmethylated spike-in A. thaliana DNA from the DNA Methylation control package (Diagenode, Seraing, Belgium).

3.2.6 cfMBD methylation capture

The DNA mixture was subjected to methylation enrichment using the MethylCap Kit (Diagenode) following the manufacture's protocol with some modifications. Total volume brought up by Buffer B was reduced from 141.8 µl to 136 µl to minimize DNA waste. The amount of MethylCap protein and magnetic beads were decreased proportionally according to the recommended input DNA to protein and beads ratio (0.2 µg protein and 3 µl beads per 100 ng DNA input). MethylCap protein was 10-fold diluted to 0.2 µg/µl using Buffer B. Single fraction elution with High Elution Buffer was applied. The eluted fraction was purified by DNA Clean & Concentrator-5 Kit. The purified DNA was divided into two parts, one for qPCR (PowerUp™ SYBR™ Green Master Mix, Thermo Fisher) amplification of spiked-in DNA for methylation enrichment quality control, another for library amplification. Recovery of the spiked-in methylated and unmethylated controls can be calculated based on cycle threshold (Ct) value of the enriched and unenriched samples. Specificity of the capture reaction can be calculated by (1 - [recovery of unmethylated control DNA over recovery of methylated control DNA]) × 100). The specificity of the reaction should be ⩾99% before proceeding to the next step.

3.2.7 DNA sequencing and alignment

Methylation-enriched DNA libraries were amplified as follows: 95 °C for 3 min, followed by 12 cycles of 98 °C for 20 s, 65 °C for 15 s and 72 °C for 30 s and a final extension of 72 °C for 1 min. During the amplification, unique indexes from primer (NEBNext Multiplex Oligos for Illumina) were added to the sequencing adapter of each sample. The amplified libraries were purified using 1 x SPRI Beads followed by a dual size selection (0.6 x followed by 1.2 x) to remove any adapter dimers. All final libraries were first quantified with the Qubit assay and qPCR-based assay using the NEBNext® Library Quant Kit for Illumina® (New England BioLabs) and then submitted to Moffitt Cancer Center Molecular Genomics Core for D1000 ScreenTape Assay for the measurement of fragment size. Libraries were sequenced on the NextSeq 550 platform (Illumina; San Diego, CA, USA), high-output 75 bp single-end read, multiplexed as 12 samples per run. After sequencing, quality control for raw sequence reads was performed using fastp (Version 0.20.1) [155] with the default settings. The sequence reads were then aligned to the human genome (hg19) using Bowtie-2 (Version 2.4.2) [166] with default settings. After the alignment, the generated sam files were converted to bam files, followed by sorting, indexing, removal of duplicate reads, and extraction of read count on chr1 - chr22 using SAMtools (Version 1.11) [216] 'view', 'sort', 'index', and 'markdup' command lines.

3.2.8 Quality control of methylation enrichment

R (Version 4.0.3 or greater) package RaMWAS (Version 1.12.0) [175] with default parameters was used for quality control of overall mapping quality and calculation of non-CpG reads percentage, non-CpG/CpG coverage (noise), and CpG density at peak. CpG annotation reference was obtained from R package annotatr (Version 1.16.0): annots='hg19_cpgs'. BEDtools (Version 2.28.0) [217] 'coverage' command line was used to call the number of sequenced reads on each CpG feature.

Read counts of each CpG feature and each sample was combined as a count matrix. Transcripts per kilobase million (TPM) normalization was performed before comparing the percentage of CpG feature read counts between different groups.

3.2.9 Differential methylation analysis of cfMBD-seq data

Rows with inter CpG regions and zero read count among all samples were filtered out from CpG feature raw count matrix. Filtered matrix were further subset for single cancer type and non-cancer control and fit a negative binomial model to call DMRs at Benjamini-Hochberg false discovery rate (BH-FDR) <0.1 (Wald test) using R package DESeq2 (Version 1.32.0) [191]. R package EnhancedVolcano (Version 1.10.0) [228] was used for visualization of fold change and BH-FDR (q value) for all CpG islands and extended CpG islands. Unsupervised hierarchical clustering was performed on Partek genomics suite (Version 7.0) for visualization of DMCGIs, using log transformed DESeq2 normalized values, z scores, Euclidean distance, and Ward Clustering. R package pcaExplorer (Version 2.18.0) [229] was used for principal component analysis of DESeq2 normalized values of top 1,000 differentially hypermethylated CpG islands (DMCGIs) selected by highest row variance. The 95% confidence ellipses for the case and control were displayed. Plasma derived DMCGIs with fold change >2 were used for intersection with tissue derived DMCGIs.

3.2.10 Methylation analyses for tumor tissue specific DMCGIs

HM450K data of primary tumors and adjacent normal tissues from patients with colon adenocarcinoma (COAD) (35 pairs), lung adenocarcinoma (LUAD) (21 pairs), and pancreatic adenocarcinoma (PAAD) (10 pairs) were acquired from TCGA. HM450K data of non-cancer individuals' PBMCs (N=61) from GEO (non-smoker controls in GSE53045) were also used to deconvolute clonal hematopoiesis effect. R package minfi (Version 1.36.0) [181] was used to call DMCs (Mean of delta beta value >0.2 and BH-FDR <0.1) between primary tumor and normal tissue / non-cancer PBMCs. R package EnhancedVolcano was used for visualization of delta beta value and q value (FDR) for all HM450K CpG sites. To make tissue derived DMCs comparable with plasma derived DMRs, all DMCs were annotated to a hg19 HM450K annotation file and their corresponding CpG islands were identified for intersection. Tissue-derived DMCGIs were identified by intersecting DMCGIs from plasma cases vs controls, primary tumors vs. normal tissues, and primary tumors vs. PBMCs. Tissue-specific DMCGIs were identified by intersecting colorectal, lung, and pancreas-derived DMCGIs. Venn diagrams were used for visualization of the intersection.

3.2.11 Machine learning analyses

Two independent cohorts were used for machine learning analyses: cfMeDIP-seq cohort and HM450K cohort. cfMeDIP-seq data of lung cancer patients (N=80) and non-cancer individuals (N=86) were used for evaluation of early cancer detection in plasma cfDNA. An independent HM450K cohort including primary tumors from TCGA (N=210 for COAD, N=385 for LUAD, and N=162 for PAAD) was used for evaluation of cancer classification performance. HM450K data were converted to a CpG islands beta value matrix by calculating the mean beta values of CpG sites annotated to the same CpG island. R package Caret (Version 6.0-88) [230] was used to partition the discovery cohort data into 100 class-balanced independent training and testing sets in an 80–20% manner. Top overlapping DMCGIs between cfMBD-seq and HM450K datasets were selected for predictive modeling analyses. R package glmnet (Version 4.1-2) [231] was used to preform regularized logistic regression model on the training sets. LASSO regularization method (alpha=1) with 10-fold cross validation was applied to determine minimum lambda penalty value. The entire process was repeated 100 times to prevent training-set biases. DMCGIs with non-zero coefficient across all repeats were determined as cancer classifiers. Performance of the predictive models was evaluated on the held-out testing set using ROC statistics. R package Rtsne (Version 0.15) [232] was used for generation of t-sne plot to visualize cancer classification in cfMBD-seq, cfMeDIP-seq, and HM450K data sets.

3.2.12. Data availability statement

R scripts and git bash used to generate the results in this study are available on GitHub (https://github.com/LiangWangLab/cfMBD-seq-clinical). The cfMBD-seq next-generation sequencing data of patient plasma samples are available upon request from the author to comply with institutional ethics regulations. Deidentified cfMBD-seq raw read count matrices for all CpG islands are available in https://www.mdpi.com/article/10.3390/cancers13225611/s1. The cfMeDIP-seq sequencing data are available upon request from the Shen et al. group [134]. The HM450K dataset is publicly available in The Cancer Genome Atlas and Gene Expression Omnibus. Primary tumor and adjacent normal tissue data can be acquired using the manifest in https://www.mdpi.com/article/10.3390/cancers13225611/s1. Peripheral blood mononuclear cell data can be found in GSE53045 (non-smoker controls).

3.3 Results

3.3.1 Significant enrichment of methylated CpG islands in cfDNA

To study the clinical feasibility of cfMBD-seq, we retrospectively profiled cfDNA methylome of 53 blood samples from patients with metastatic carcinoma of the colon/rectum (N=13), lung (N=12), and pancreas (N=12), and from cancer-free individuals (N=16). We quantified cfDNA concentration from plasma samples and showed that cancer patients had higher cfDNA yield than non-cancer controls (**Figure 17a**). To investigate methylation capture efficiency of cfMBD-seq, we compared spiked-in controls between methylated and unmethylated *A. thaliana* DNA in the capture reaction and observed a median specificity at 99.3% [99.16% (Q1) - 99.43% (Q3)] across all samples (**Figure 17b**). From sequencing data, we filtered out duplicate reads and reads with low alignment scores from total sequence reads (41.62 [38.75 - 44.43] million) and obtained 35.33 [32.77 - 37.37] million high-quality reads (**Figure 17c**). We then investigated genome-wide methylation enrichment and found that the number of captured fragments without any CpG tandem accounted for only 1.47% [1.33% - 1.59%] of high-quality reads (**Figure 17d**). The coverage ratio of fragments without any CpG tandem to fragments with at least one CpG, known as noise, was 0.15 [0.13 - 0.17] (**Figure 17e**). The median CpG density of fragments with the highest read coverage was 25.2 [24.2 - 25.7] (**Figure 17f**), corresponding to high-CpG density regions - CpG islands. Intrigued by the high sequencing coverage on CpG islands, we further studied the distribution of sequence reads by calculating the percentage of normalized reads on different CpG annotation features (i.e., CpG islands, CpG shores, CpG shelves, and inter CpG regions). We found a median of 42.16% [39.47 - 45.15] of reads mapped to CpG islands, when CpG islands only account for 0.7% of the hg19 reference genome (**Figure 18a&d**). Since methylation alterations may occur at a short distance away from the CpG islands [218], we also calculated the sum of reads mapped to extended CpG islands (i.e., CpG islands, CpG shores, and CpG shelves). A median of 91.46% [90.89% - 92.13%] of reads were mapped to the extended CpG islands, which accounts for only 6.72% of the reference genome (**Figure 18b-d**). These results demonstrate that most of the sequence reads captured by cfMBD-seq were significantly enriched on CpG island-centered regions, illustrating successful cfMBD-seq methylation enrichment and library construction across all samples.

3.3.2 Differential methylation analyses between cases and controls

To identify differences in methylation patterns between cases and controls, we generated a read count matrix for each cancer type versus non-cancer control. In this matrix, each row represents a different CpG feature, and each column represents a unique individual sample. We then removed rows annotated as inter CpG and rows with 0 read count across all samples and obtained 115459 genomic regions. Next, we performed differential methylation analysis based on a negative binomial model of feature counts at a significance level of 0.1 using Benjamini-Hochberg false discovery rate (BH-FDR) and identified 2722, 3033, and 2831 DMRs for colorectal, lung, and pancreatic cancer, respectively (**Figure 19a-c**). We further filtered these DMRs using a more stringent criteria: absolute fold change >2, which resulted in 2009 DMRs (2007 hypermethylated and 2 hypomethylated) in colorectal cancer, 1818 DMRs (1814 hypermethylated and 4 hypomethylated) in lung cancer, and 1488 DMRs (1482 hypermethylated and 6 hypomethylated) in pancreatic cancer. As the majority of the remaining DMRs were hypermethylated, and most of them were CpG islands (97%, 85%, and 93% in colorectal, lung and pancreatic cancer patients, respectively), to enhance computational efficiency, we reduced our dataset to 26441 CpG islands and applied the same criteria for differential methylation analysis (BH-FDR<0.1 and fold change >2). This optimized analysis identified 1759, 1783, and 1548 differentially hypermethylated CpG islands (DMCGIs) in colorectal, lung, and pancreatic cancer, respectively (**Figure 19d-f**). Unsupervised hierarchical clustering of the top 100 hypermethylated CpG islands ranked by p-value well distinguished cancer patients from non-cancer individuals by dividing these groups into two clusters (**Figure 20**). Principal component analysis (PCA) using the top 1000 DMCGIs revealed partitioning of cancer patients from the non-cancer controls (**Figure 21**). In the PCA plots, non-cancer samples clustered tightly together, while cancer samples were not clustered, which may be attributed to tumor heterogeneity. These combined findings suggest that cfMBD-seq can reliably identify DMCGIs between the plasma cfDNA of cancer patients and non-cancer controls.

3.3.3 Overlap between tissue-derived and cfDNA-derived DMCGIs

To explore whether the DMCGIs detected by cfMBD-seq were originated from tumor tissues, we acquired Infinium HumanMethylation450 BeadChip (HM450K) data of primary tumors and matched adjacent normal tissues from the same patients, including colon adenocarcinoma (COAD, 35 pairs), lung adenocarcinoma (LUAD, 21 pairs), and pancreatic adenocarcinoma (PAAD, 10 pairs) from The Cancer Genome Atlas (TCGA) (**Figure 22a**). We identified 21274, 7635, and 7458 hypermethylated differentially methylated CpG sites (DMCs) (Mean of Δbeta value >0.2, BH-FDR<0.1, F-test) between primary tumors and matched normal tissues of COAD, LUAD, and PADD, respectively (**Figure 22b-d**). To make HM450K results comparable to cfMBD-seq, we excluded the DMCs that were not annotated to CpG islands and kept the remaining 94.05%, 84.44%, and 90.73% of DMCs in the three cancer types. After further removal of duplicated CpG islands, we obtained 4630, 2588, 2478 unique DMCGIs for COAD, LUAD, and PAAD, respectively. As non-tumor-derived cfDNA is mostly released from peripheral blood mononuclear cells (PBMCs), we conducted an analysis to determine whether the DMCGIs identified by cfMBD-seq were not derived from clonal hematopoiesis differences between cases and controls. For this purpose, we performed similar differential methylation analyses between HM450K data of primary tumors and cancer-free individuals' PBMCs (N=61 from Gene Expression Omnibus (GEO), non-smoker controls in GSE53045) and identified a set of DMCs for each cancer type (**Figure 22e-g**). After annotation and exclusion of DMCs, we obtained 7838, 4906, and 5613 unique DMCGIs for COAD, LUAD, and PAAD, respectively. Intersection analyses of three sets of DMCGIs showed that 84.5% of colorectal (1486/1759), 52.7% of lung (939/1783), and 57.9% of pancreatic (896/1548) cancer DMCGIs detected by cfMBD-seq overlapped with not only DMCGIs between primary tumor and adjacent normal tissue, but also DMCGIs between primary tumor and PBMCs (**Figure 22h**). These findings suggest that plasma derived DMCGIs detected by cfMBD-seq were mainly driven by tumor-specific DNA methylation patterns rather than by background noise of cell composition in the tumor microenvironment.

3.3.4 Differentially methylated CpG islands for lung cancer detection

Since most HM450K data are originated from early-stage cancer tumor tissue samples, we hypothesized that the identified overlapping DMCGIs can be used for the early cancer detection. To test this hypothesis, we acquired an additional cohort of 166 plasma samples including 80 lung cancer patients (N=22 with early-stage disease) and 86 non-cancer individuals from a previous cfMeDIP-seq study [134] (**Figure 23a**). t-distributed stochastic neighbor embedding (t-sne) plot using the 939 overlapping lung cancer DMCGIs identified a clear separation between lung cancer and non-cancer individuals in the cfMeDIP-seq cohort, and only 5 individuals were misclassified (**Figure 23b**). To rigorously evaluate the utility of these overlapping DMCGIs for cancer detection, we selected the top 300 lung cancer DMCGIs based on their rank on fold change in the cfMBD-seq results and carried out a set of machine learning analyses on the cfMeDIP-seq cohort. We randomly split these samples into balanced training (80%) and testing (20%) sets. To select the most discriminating markers, we trained a series of case-versus-control binomial generalized linear models (logistic regression) with least absolute shrinkage and selection operator (LASSO) regularization using these top features on the training sets. The process was repeated 100 times to prevent training-set biases. Eventually, we identified 3 DMCGIs (chr1:243646395-243646888, chr8:99985734-99986983, and chr21:38068194-38073891) that had non-zero coefficients across all repeats and selected those as cancer classifier. The normalized sequence reads of the 3 DMCGIs were significantly higher in lung cancer patients than in non-cancer individuals (**Figure 23c**). To evaluate the performance of the classifier, we fit the predictive model on the testing dataset and used receive operating characteristic (ROC) statistics to calculate area under the ROC curve (AUC) for evaluation. The results showed that the model can predict lung cancer in the testing set with high accuracy (AUC=0.949 [0.929-0.982]) (**Figure 23d**). Using only the 3 DMCGIs for t-sne plot, all samples were correctly classified (**Figure 23e**). These results suggest that early cancer detection is possible when using tissue-specific DMCGIs identified by cfMBD-seq.

3.3.5 Differentially methylated CpG islands for cancer classification

To further investigate the candidate DMCGIs shared between cfDNA and tumor tissue, we intersected the three sets of selected DMCGIs for colorectal (N=1486), lung (N=939), and pancreatic (N=896) cancer. We identified a total of 1271 cancer type specific DMCGIs, including 738 for colorectal cancer, 370 for lung cancer, and 163 for pancreatic cancer. Also, a total of 266 DMCGIs were shared by these three cancer types (**Figure 24a**). To rigorously evaluate the performance of these cancer-type specific DMCGIs in cancer classification, we acquired an additional independent TCGA HM450K data cohort including primary tumors for COAD (N=210), LUAD (N=385), and PAAD (N=162) (**Figure 24b**). To convert HM450K data to CpG islands-based beta value, we filtered out CpG sites that weren't annotated to CpG islands from 485577 HM450K locus and used the remaining 309465 CpG sites for subsequent analysis. Given the methylation level between neighboring CpG sites are positively correlated, we calculated the mean beta values of CpG sites annotated to the same CpG island and generated a beta value matrix for all CpG islands. We then performed similar machine learning analyses on the HM450K cohort using the top 100 cancer type-specific DMCGIs. The analyses consisted of 4:1 sample partition, LASSO regularization, and logistic regression modeling. Rather than a case-versus-control model, here we built a one-versus-all-others model for each cancer type. After 100 repeats of the training process, we identified 3 colorectal, 16 lung, 6 pancreatic specific DMCGIs (non-zero coefficients) as cancer type classifier. Again, we fit the predictive model on the held-out testing set and applied ROC statistics for evaluation. The results showed great performance in the prediction of cancer type (median AUC=1 for COAD, 1 for LUAD, and 0.989 for PAAD). Methylation levels of cancer type specific DMCGIs are higher in its specific cancer type than in other cancer types (**Figure 24c**). To better visualize the classification performance, we generated t-sne plot using these cancer type classifiers and observed clear separation by tumor type in the cfMBD-seq plasma cohort (**Figure 24d**). This separation was notably reproduced in the HM450K cohort of 757 cancer tissue and 61 blood cell samples (**Figure 24e**). These results indicate the robust ability of cfMBD-seq to recover tumor tissue-derived methylation profiles in cfDNA across a range of cancer types and enable cancer type classification.

3.3.6 Gene annotation of differentially methylated CpG islands

To gain an understanding of the biological process behind cancer type specific DMCGIs, we linked these DMCGIs to their associated genes (**Table 4**). Some DMCGIs were annotated to gene promoter regions. We found that several genes with promoter CpG island hypermethylation are implicated in immune response, which is generally downregulated in cancer [233]. For example, the protein encoded by *PTGER4* is a member of the G-protein coupled receptor family that can activate T-cell factor signaling [234]. We not only identified DMCGIs in gene promoter regions, but also found DMCGIs in gene bodies and intergenic regions. (**Table 4**). In contrast to promoter CpG islands hypermethylation that prevents gene expression, hypermethylation in gene body CpG islands can enhance gene expression levels [235]. Consistent with our findings, several genes with gene body hypermethylation were associated with the regulation of developmental processes. For example, the protein encoded by *WNT6* and *HOXB8* has been implicated in oncogenesis and in several developmental processes such as embryogenesis. Overexpression of both *WNT6* and

HOXB8 plays key roles in carcinogenesis [236,237]. These results suggest that cfMBD-seq can capture tumor relevant biological signals in the plasma cfDNA methylome. Taken together, our results indicate that DMCGIs in cfDNA are useful in cancer detection and classification, suggesting that tumor-derived epigenomic signals are retained in the cfDNA methylome profiled by cfMBD-seq.

3.4 Discussion

Blood-based assays that can identify the tissue of origin associated with cfDNA fragments could be instrumental in detecting and classifying malignancies based on histological subtypes. In this study, we highlight the potential of hypermethylated CpG islands in cancer detection and classification. Generally, sequencing data from methylation enrichment-based methods are analyzed by comparing the relative abundance of captured fragments. The genome is divided into non-overlapping adjacent genomic windows of a specified width and the number of sequencing read counts is called for each window. Taking 300 bp window as an example, there will be more than 10 million genomic regions which requires a significant amount of computing memory. In this study, instead of genomic windows, we called read counts according to CpG annotation features. This is because MBD methylation enrichment has bias toward hypermethylation on high CpG density regions [140]. We found that 42.16% of the sequence reads in this study were mapped to CpG islands, and that 91.46% of the reads were mapped to the extended CpG islands, which account for only a small fraction of the human genome (**Figure 18**). Therefore, by excluding the large fraction of low value inter CpG regions, the computational efficiency was significantly enhanced. Additionally, well established RNA-seq data analysis packages such as DESeq2 can be directly applied to the CpG features read count matrix. Together, this CpG island-centered strategy is a preferred data analysis method for cfMBD-seq.

Differential methylation analysis based on a negative binomial model of CpG island read counts identified overwhelming differentially hypermethylated CpG islands (**Figure 19**). This is consistent with the fact that the tumor methylome is characterized by DNA methylation alterations with CpG islands-specific hypermethylation. Unlike genomic DNA from primary tumor tissue that can perfectly discriminate from non-cancer specimens, cfDNA in blood has much lower tumor-derived signal and much higher confounding signals from normal cells. Additionally, pre-analytical factors such as plasma collection and cfDNA library preparation can also affect the identification of methylation signatures. These factors may partially explain why both clustering and principal component analysis didn't perfectly segregate cancer and non-cancer specimens (**Figure 21**). In this study, confounding factors such as age and gender were not well matched between the cases and controls, which may result in false positive DMCGIs. To assess whether the DMCGIs identified by cfMBD-seq represented tumor-derived DNA methylation changes, we compared our findings against the HM450K primary tumor tissue data. We first identified a set of DMCGIs between paired primary tumor tissues and adjacent normal tissues. Since non-tumor-derived cfDNA released from blood cells can also lead to false positive results, we then identified a set of DMCGIs between primary tumor tissues and non-cancer PBMCs to deconvolute the clonal hematopoiesis effect. In our intersection analysis, the majority of the DMCGIs identified in plasma using cfMBD-seq were consistent with tumor tissue-derived DMCGIs across all analyzed cancer types (**Figure 22h**).

The main limitation of this study is the small sample size which prevented us from building prediction models using cfMBD-seq dataset. Instead, we selected to use the cfMeDIP-seq and HM450K datasets for predictive modeling. In the LASSO regularized logistic regression analysis using overlapping lung cancer DMCGIs in cfMeDIP-seq dataset, the model was able to discriminate lung cancer patients and non-cancer controls in the testing set with high accuracy (**Figure 23d**). However, when we tried fitting the model to our cfMBD-seq dataset for validation purpose, the prediction performance was relatively poor (data not shown). Although the methylation capture principle and data analysis pipelines of these two technologies are highly similar, the capture efficiencies on fragments with different CpG density are different. cfMeDIP-seq preferentially enriches methylated regions with a modest CpG density, while cfMBD-seq captures a broad range of CpG densities and identifies a larger proportion of CpG islands [101]. These differences may explain the inferior performance of the classifier in our study cohort. Additionally, it is important to note that HM450K and cfMBD-seq are completely different technological platforms. Unlike bisulfite conversion-based HM450K, cfMBD-seq is an enrichment-based method that cannot provide the absolute methylation level at each CpG site. Taking advantage of the fact that the methylation level between neighboring CpG sites is positively correlated, we transformed the CpG sites beta value matrix into a CpG islands beta value matrix. This transformation not only mitigates the disadvantage that HM450K has poor coverage of all CpG sites, but also makes HM450K data comparable with cfMBD-seq DMCGIs. However, since HM450K data are derived from tumor tissue genomic DNA, cancer type classifiers identified from the predictive models cannot be directly applied for cancer classification on plasma cfDNA-based methylation data. Future studies with larger patient cohorts are needed to validate our findings.

The potential clinical application of cfMBD-seq is worth discussing. Currently, PCR-based assays such as Epi proColon have been approved by FDA for clinical detection of cancer. However, PCR-based assays are restricted by their low throughput nature and thus is only suitable for the detection of a specific cancer type. On the other hand, sequencing-based assays are more feasible for pan-cancer screening. GRAIL takes advantages of targeted bisulfite sequencing by combining methylation signatures of different cancer types and achieves high accuracy in multi-cancer early detection. Adela is a recently established company that is focused on the detection of cancer by means of cfMeDIP-seq, another enrichment-based cfDNA methylation profiling method. Adela debuts with $60 million in Series A financing and is ready to translate to the clinic. The success of Adela gives promise to cfMBD-seq since it outperforms cfMeDIP-seq in many aspects. cfMBD-seq is capable of efficiently capturing and preferentially targeting the methylated CGIs from the entire methylome using ultra-low input of plasma cfDNA. Unlike targeted bisulfite sequencing, this enrichment-based profiling method shows robust cancer detection and classification performance without the need of complicated primer and probe design procedures for the bisulfite-converted sites. For the detection of a specific cancer type, although cfMBD-seq is cost-efficient among sequencing-based assays, it is still less competitive in price compared to PCR-based assays. Nonetheless, for a minimally invasive early cancer detection test applied in the general populations, pan-cancer screening is preferred. Overall, cfMBD-seq has a promising future in translation to clinic, but reliable classifiers for more cancer types remain to be identified.

In this proof of principle study, we provide important insights into the possible future clinical applications of cfMBD-seq. Highlights of the study include: 1) cfMBD-seq enables the identification of cancer-associated DMCGIs from plasma cfDNA in cancer patients; 2) the identified DMCGIs are mainly driven by tumor-specific DNA methylation patterns and demonstrate promise for future studies using this technology for cancer detection and classification; 3) the most discriminating DMCGIs selected by our prediction models are associated with important biological processes that are contribute to carcinogenesis. In summary, cfMBD-seq is a non-invasive, cost-effective, bisulfite-free, and sensitive methylation profiling method for capturing of hypermethylated CpG islands in cfDNA. This study demonstrates potential clinical feasibility of cfMBD-seq. The current results provide considerably strong justification for future biomarker discovery and validation in large-scale patient populations. Our findings underscore the utility of differentially hypermethylated CpG islands in cfDNA for accurate cancer detection and multi-cancer classification.

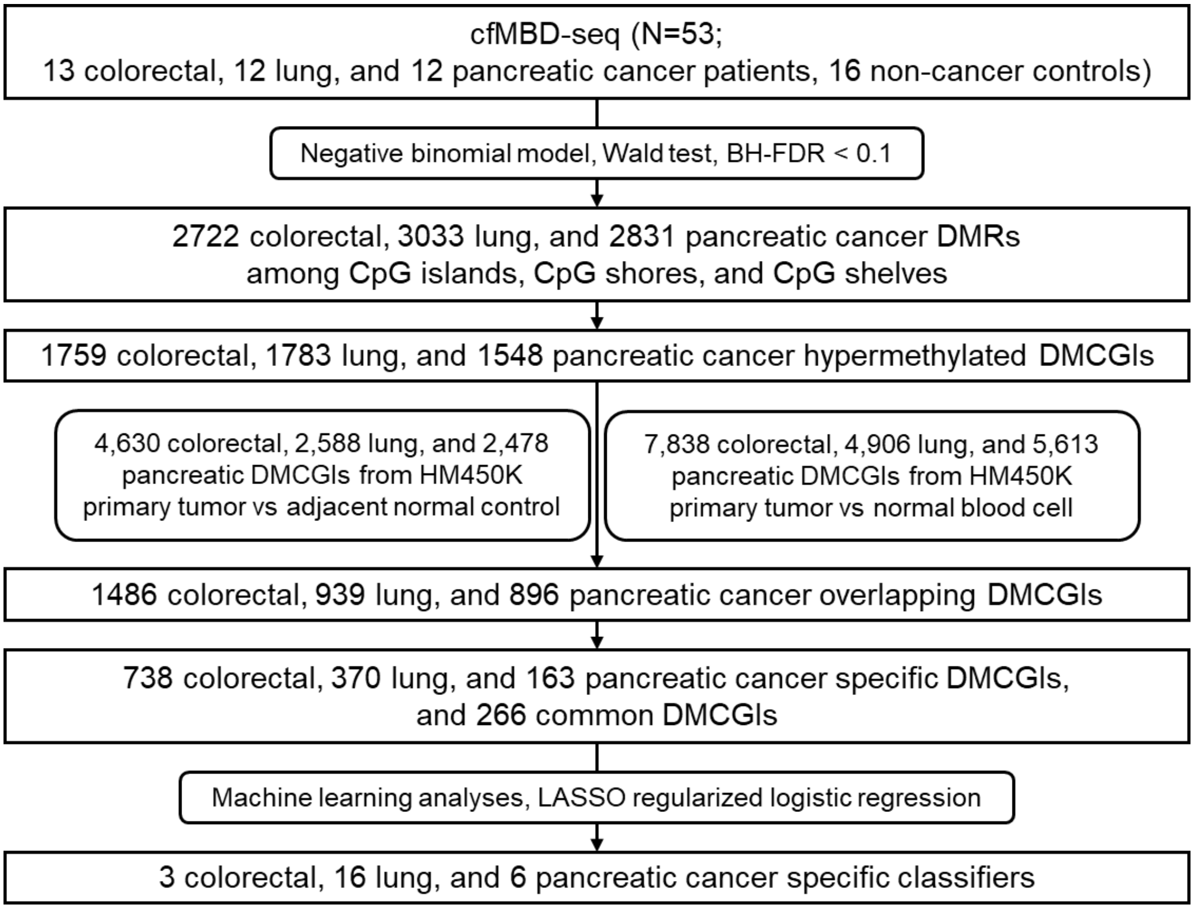

cfMBD-seq (N=53;
13 colorectal, 12 lung, and 12 pancreatic cancer patients, 16 non-cancer controls)
Negative binomial model, Wald test, BH-FDR < 0.1
2722 colorectal, 3033 lung, and 2831 pancreatic cancer DMRs
among CpG islands, CpG shores, and CpG shelves
1759 colorectal, 1783 lung, and 1548 pancreatic cancer hypermethylated DMCGIs
4,630 colorectal, 2,588 lung, and 2,478 pancreatic DMCGIs from HM450K primary tumor vs adjacent normal control
7,838 colorectal, 4,906 lung, and 5,613 pancreatic DMCGIs from HM450K primary tumor vs normal blood cell
1486 colorectal, 939 lung, and 896 pancreatic cancer overlapping DMCGIs
738 colorectal, 370 lung, and 163 pancreatic cancer specific DMCGIs,
and 266 common DMCGIs
Machine learning analyses, LASSO regularized logistic regression
3 colorectal, 16 lung, and 6 pancreatic cancer specific classifiers

Figure 16. Workflow chart of data generation and analysis

BH-FDR, Benjamini-Hochberg false discovery rate; DMRs, differentially methylated regions; DMCGIs, differentially methylated CpG islands; LASSO, least absolute shrinkage and selection operator.

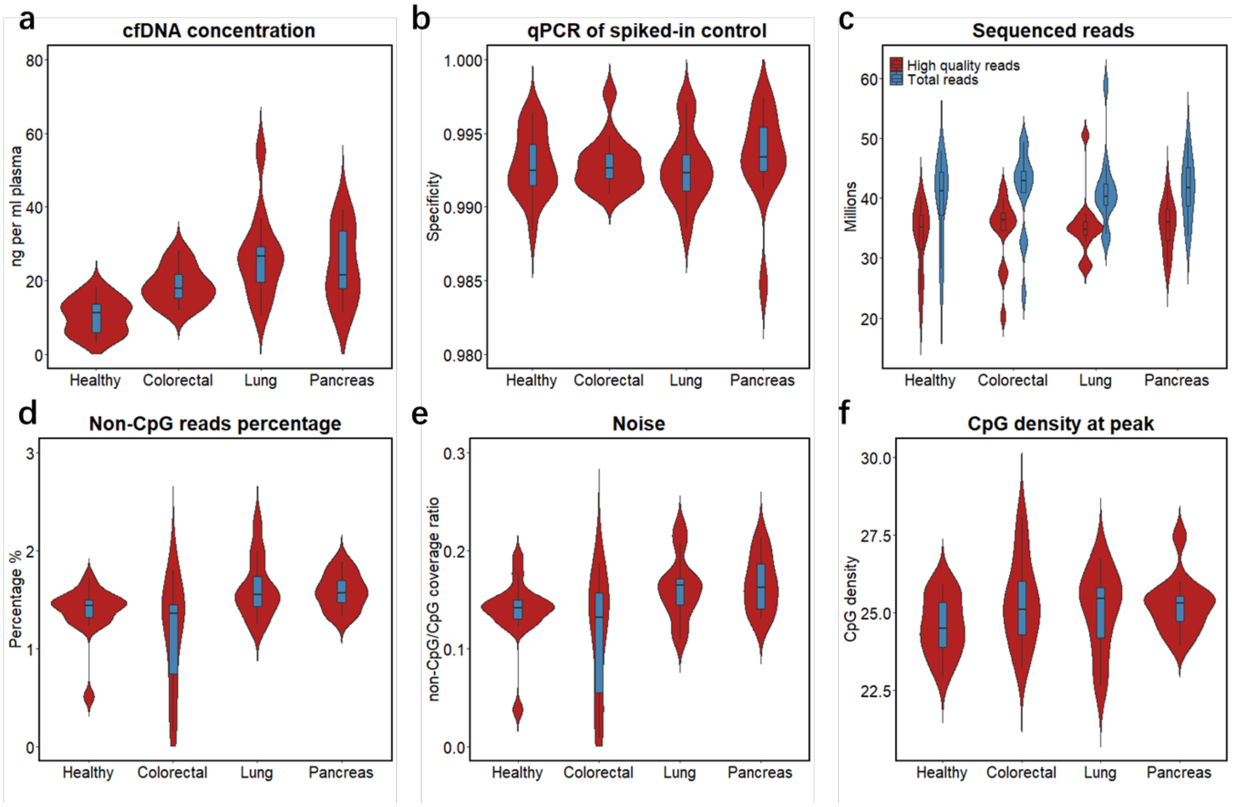


Figure 17. Quality controls of cfMBD-seq

a. cfDNA concentration (ng cfDNA per ml plasma) from colorectal cancer (N=13), lung cancer (N=12), pancreatic cancer (N=12) patients, and non-cancer controls (N=16). b. Specificity of MBD methylation capture reaction across different groups (i.e., Healthy, non-cancer individuals; Colorectal, colorectal cancer patients; Lung, lung cancer patients; Pancreas, pancreatic cancer patients) calculated using qPCR Ct value of meth-ylated and unmethylated spiked-in A. thaliana DNA. c. Total sequence reads and high-quality sequence reads across different groups. d. Percentage of sequence reads that doesn't contain any CpG tandem across different groups. e. Ratio of average non-CpG coverage to average CpG coverage across different groups. Non-CpG coverage is defined as the average coverage of fragments without any CpG tandem. CpG coverage is defined as the average coverage of fragments with no less than one CpG tandem. f. CpG density at peak across different groups. CpG density is defined as number of CpG tandems per fragment. Peak is defined as fragments with highest coverage.

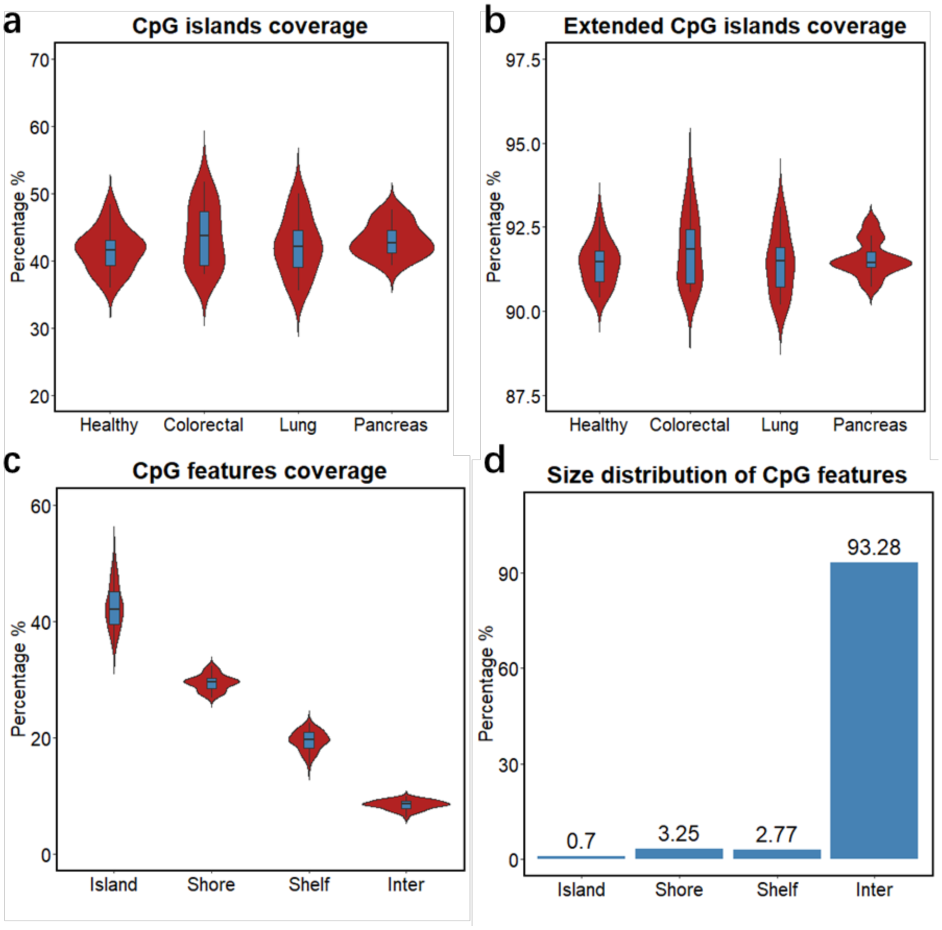


Figure 18. Quality controls of CpG islands-centered enrichment

a. Percentage of transcripts per million (TPM) normalized reads on CpG islands across different groups. b. Percentage of TPM normalized reads on CpG islands/shores/shelves across different groups. For all box plots, the extremes of the boxes represent the upper and lower quartiles, and the center lines define the median. Whiskers indicate 1.5x interquartile range. c. Percentage of sequencing coverage across different CpG annotation features (i.e., CpG islands, CpG shores, CpG shelves, and inter CpG regions) for all samples. d. Percentage of different CpG annotation features in base pair size in hg19 human genome. For all box plots, the extremes of the boxes represent the upper and lower quartiles, and the center lines define the median. Whiskers indicate 1.5x interquartile range.

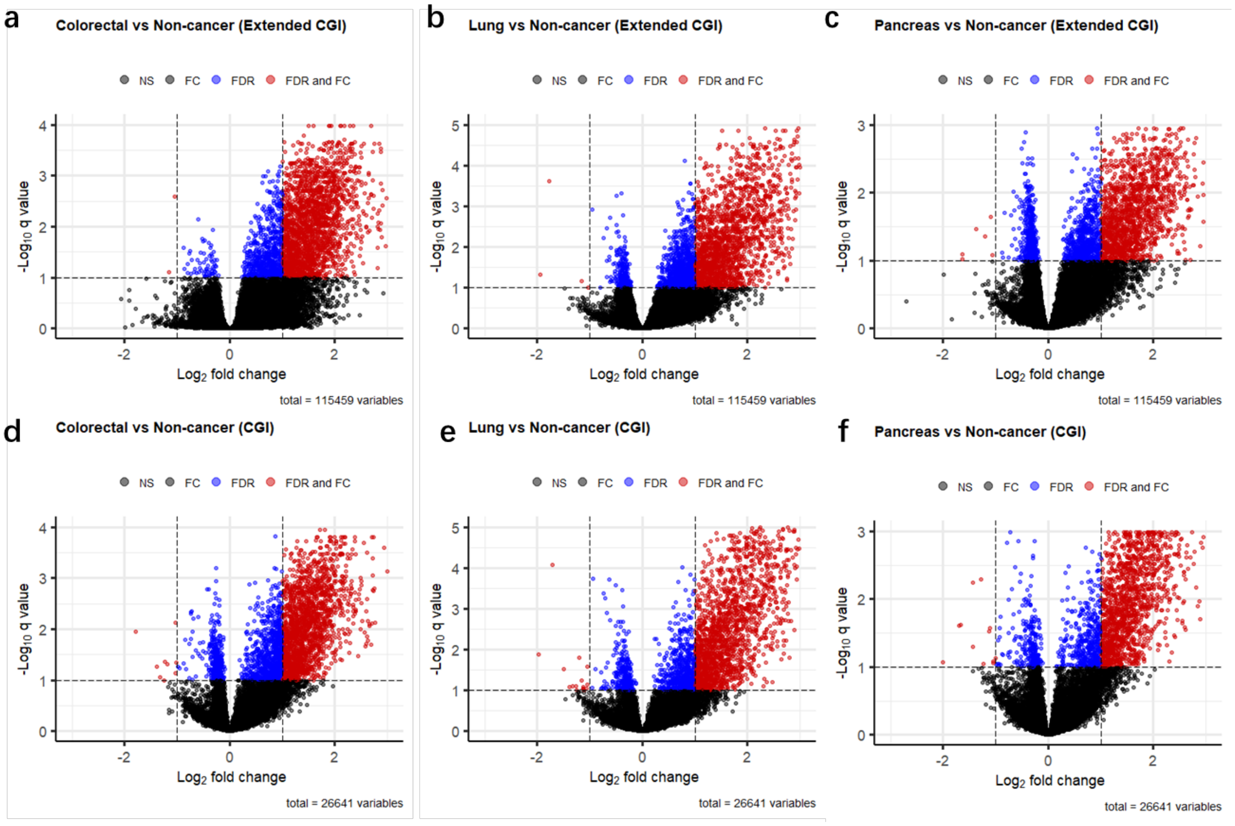


Figure 19. Differentially methylated regions between cases and controls

a-c. Volcano plots of differentially methylated regions (DMRs) at extended CpG islands (CGI) (i.e., CpG islands, CpG shores, and CpG shelves) between colorectal cancer (N=13) (a) / lung cancer patients (N=12) (b) pancreatic cancer (N=12) (c) and non-cancer controls (N=16). Black dots indicate non-significant regions. Blue and red dots indicate statistical significance at Benjamini-Hochberg false discovery rate (FDR) < 0.1 (negative binomial model, Wald test). Red dots also indicate regions with absolute fold change (FC) >2. d-f. Volcano plots of DMRs at CpG islands between colorectal cancer (d) / lung cancer patients (e) pancreatic cancer (f) patients and non-cancer controls.

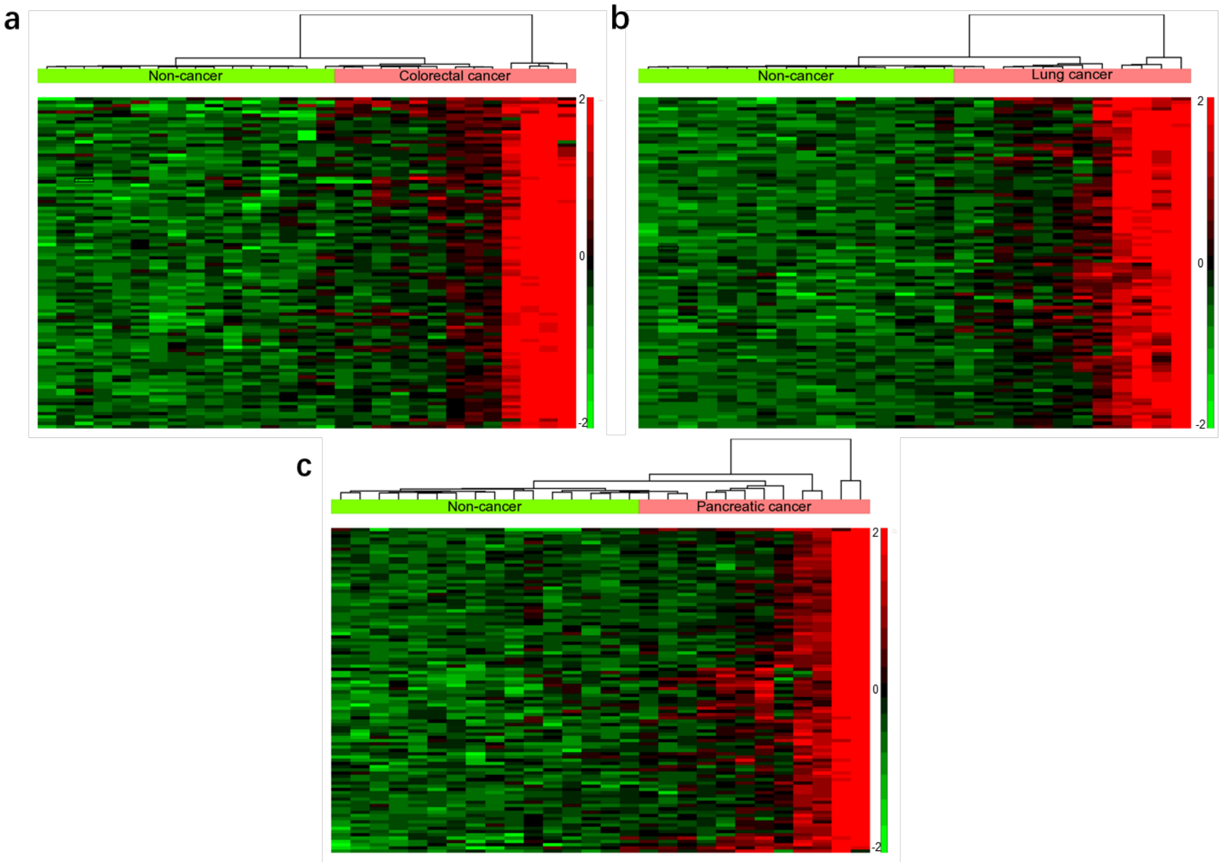


Figure 20. Heatmap of cfMBD-seq DMRs between cases and controls

a-c. Unsupervised hierarchical clustering (z scores normalization of DESeq2 normalized counts, Euclidean distance, and Ward Clustering) of the top 100 differentially hypermethylated CpG islands between colorectal cancer (a) / lung cancer (b) / pancreatic cancer (c) patients and non-cancer controls. Dendrogram shows separation by sample type (case or control).

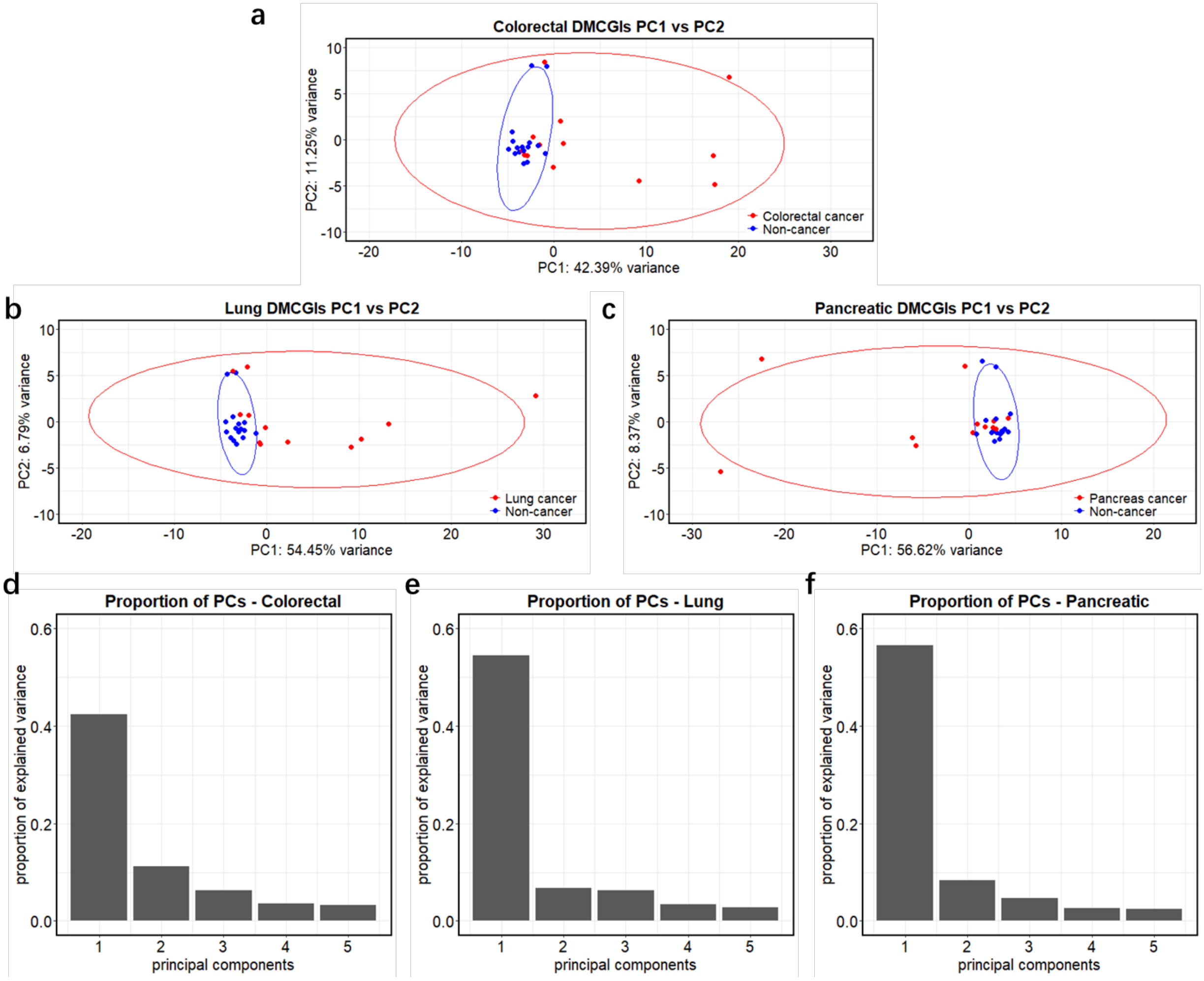


Figure 21. Principal component analysis of DMRs detected by cfMBD-seq

a-c. Principal component analysis using DESeq2 normalized counts of top 1,000 differentially hypermethylated CpG islands between colorectal cancer (a) / lung cancer (b) / pancreatic cancer (c) patients and non-cancer controls. The 95% confidence ellipses for the case and control are displayed. d-f. Proportion of variance explained by each principal component.

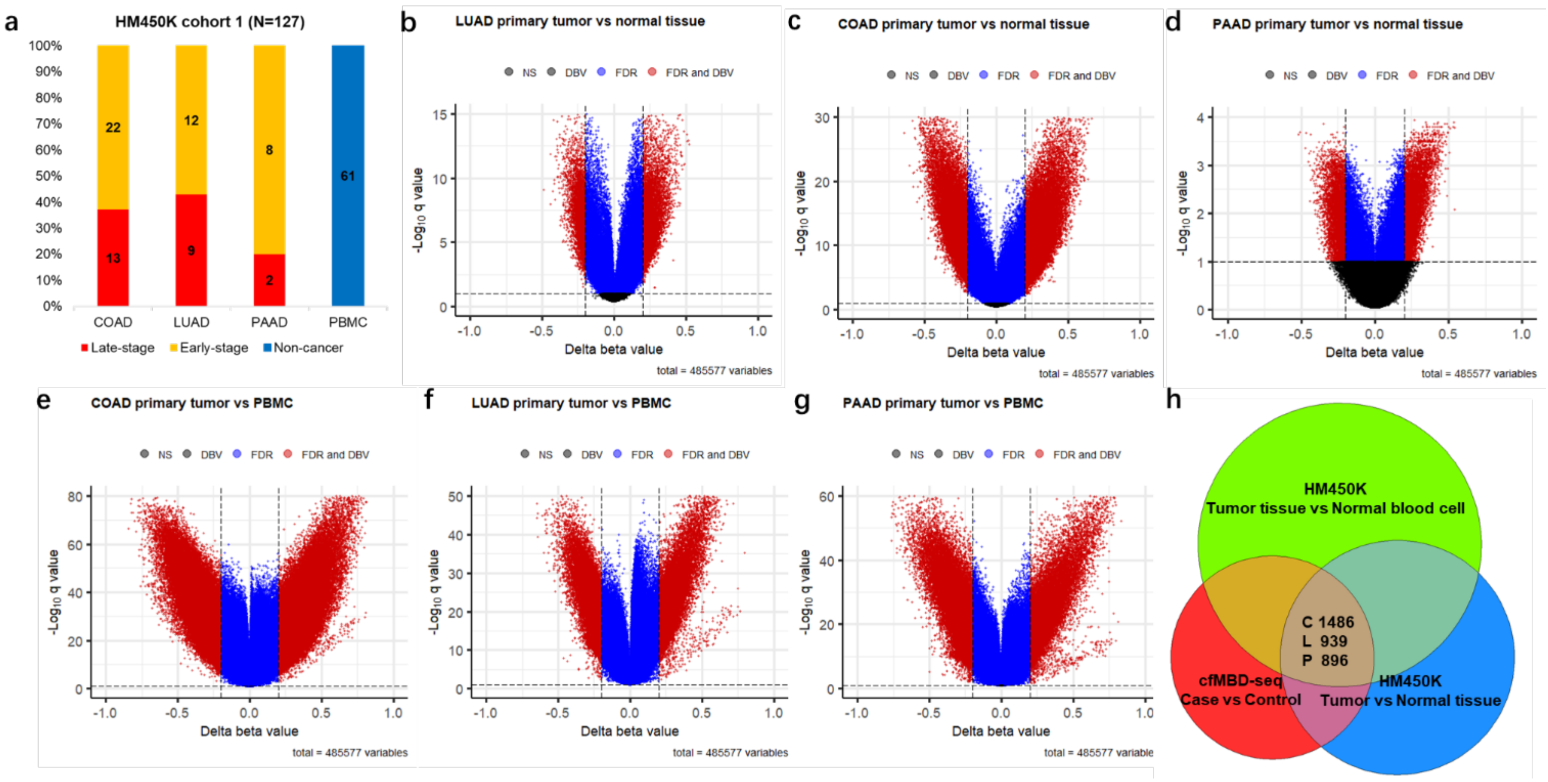


Figure 22. HM450K DMCs between primary tumors and normal tissues / blood cells

a. Pathology stage (according to the AJCC/UICC 7th Edition) in the HM450K cohort including N=66 paired primary tumors and adjacent normal tissues, and N=61 non-cancer peripheral blood mononuclear cells (PBMCs). Early-stage consists of stage I and II. Late-stage consists of stage III and IV. b-d. Volcano plots of DMCs between primary tumors and adjacent normal tissues for COAD (N=35) (b) LUAD (N=21) (c) or PAAD (N=10) (d) from HM450K data. e-g. Volcano plots of DMCs between primary tumors and PBMCs for COAD (e), LUAD (N=21) (f), or PAAD (g). For all volcano plots, black dots indicate non-significant regions. Blue and red dots indicate regions significant at Benjamini-Hochberg false discovery rate (BH-FDR) < 0.1 (F-test). Red dots also indicate regions with mean of □beta value >0.2. h. Venn diagram showing the number of overlapping regions between plasma-derived differentially methylated CpG islands (DMCGIs) from cfMBD-seq and tissues-derived DMCGIs from HM450K in three cancer types (i.e., C, colorectal cancer; L, lung cancer; P, pancreatic cancer.

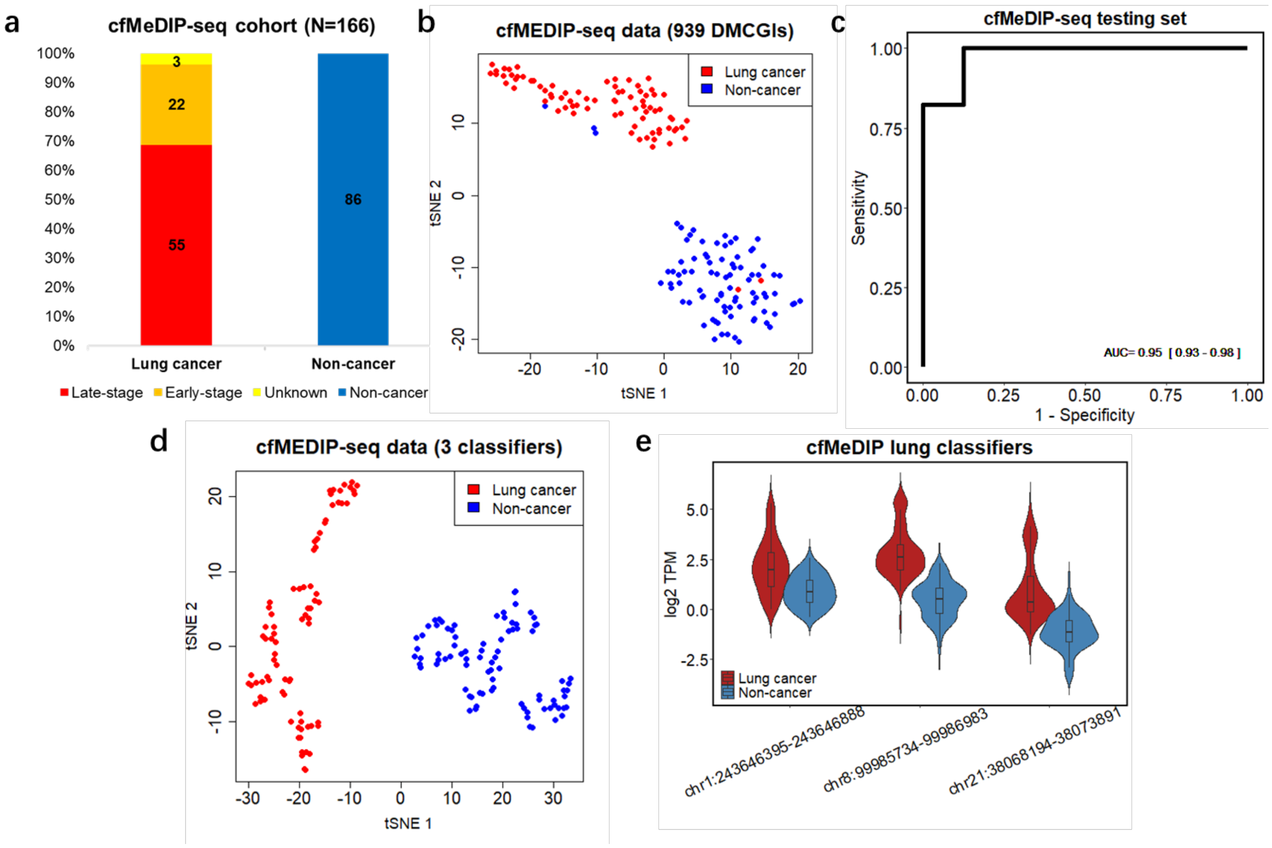


Figure 23. Differentially methylated CpG islands in lung cancer detection

a. Pathology stage (according to the AJCC/UICC 7th Edition) in the cfMeDIP-seq cohort. Early-stage consists of stage I and II. Late-stage consists of stage III and IV. b. t-distributed stochastic neighbor embedding (t-sne) plot using all 939 lung DMCGIs that are overlapped between cfMBD-seq and HM450K data for the entire cfMeDIP-seq plasma samples (N=166). c. Predictive modeling using LASSO regularized logistic regression case-versus-control models on cfMeDIP-seq cohort including lung cancer patients (N=80) and non-cancer controls (N=86). ROC curve for 20% of held-out testing set is shown. AUC values represent median and interquartile range for 100 repeats of the model. d. t-sne plot using the top 3 lung cancer specific DMCGIs identified from training set for plasma samples of the entire cfMeDIP-seq cohort. e. Log transformed transcripts per kilobase million (TPM) sequence reads of the classifier from the cfMeDIP-seq training set. The extremes of the boxes define the upper and lower quartiles, and the center lines define the median. Whiskers indicate 1.5x interquartile range.

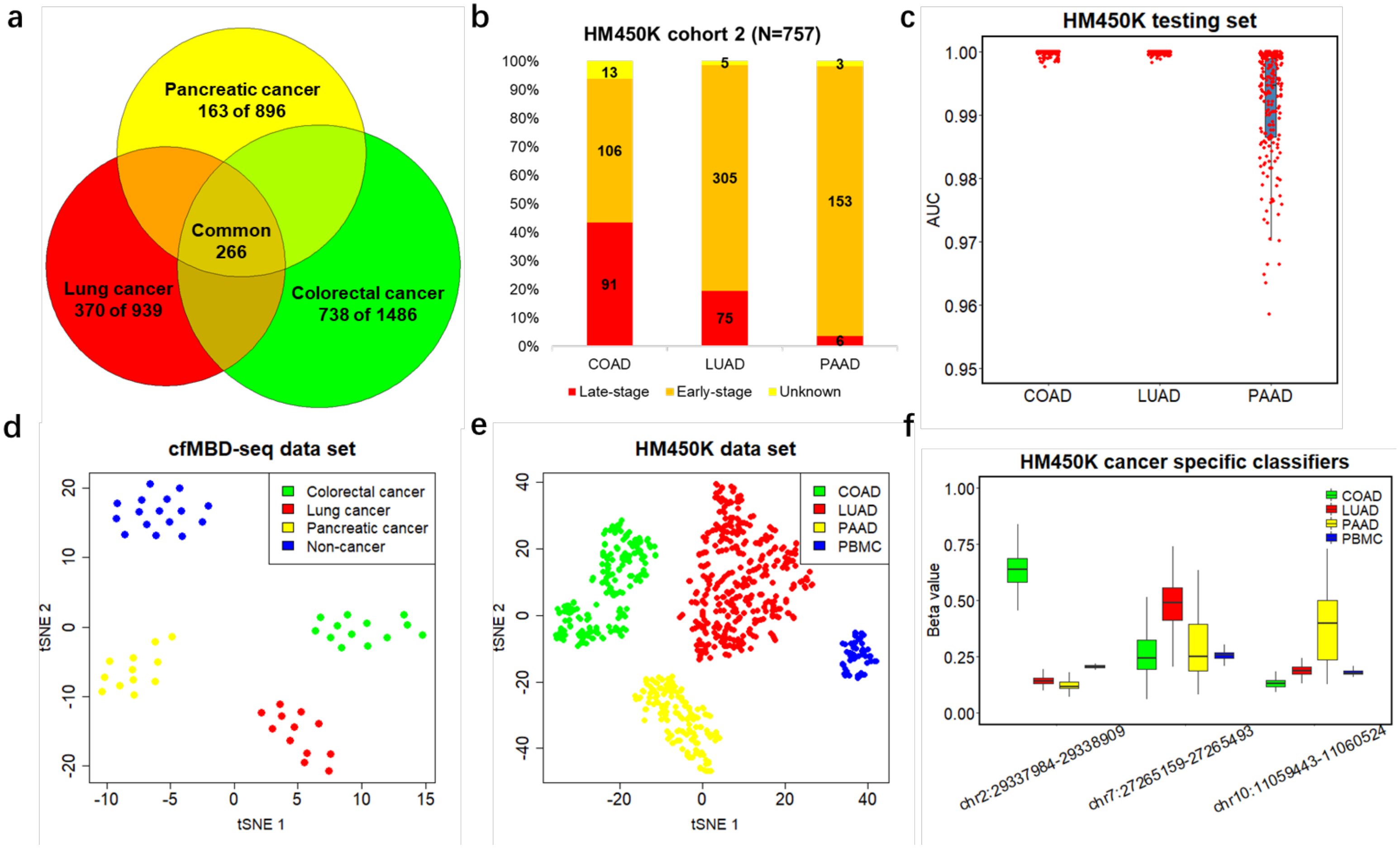


Figure 24. Differentially methylated CpG islands in cancer classification

a. Venn diagram showing the number of tissue specific DMCGIs for each cancer type and the number of DMCGIs that are common in all three cancer types. b. Pathology stage (according to the AJCC/UICC 7th Edition) in the TCGA HM450K cohort of different tumor (N=757). Early-stage consists of stage I and II. Late-stage consists of stage III and IV. c. Predictive modeling using LASSO regularized logistic regression one-versus-all-others models on the HM450K cohort including 210 colon adenocarcinoma (COAD) samples, 385 lung adenocarcinoma (LUAD) samples, and 162 pan-creatic adenocarcinoma (PAAD) samples. Area under the curve (AUC) values are

Figure 24. (Continued) calculated from 20% of held-out testing set. Boxplots represent median and interquartile range for 100 repeats of the models. d & e. t-sne plot using cancer type specific classifiers identified from training set for the entire cfMBD-seq plasma cohort (N=53) (d) and HM450K tissue cohort (N=757 primary tumor and N=61 non-cancer PBMCs) (e). f. Beta value of cancer type specific classifiers (Colorectal cancer specific: chr2:29337984-29338909; Lung cancer specific: chr7:27265159-27265493; Pancreatic cancer specific: chr10:11059443-11060524) across COAD, LUAD, PAAD, and PBMC samples. The extremes of the boxes define the upper and lower quartiles, and the center lines define the median. Whiskers indicate 1.5x interquartile range.

Table 3 Clinical demographic characteristics of patients in the cfMBD-seq cohort

Type		Non-cancer	Colorectal cancer	Lung cancer	Pancreatic cancer
Age	Mean	49.38	67.88	58.33	61.25
	Median	52.5	72.5	57.5	62.5
	Range	42.5-52.5	47.5-82.5	42.5-67.5	42.5-67.5
Gender	Male	5	11	9	8
	Female	11	2	3	4
Race	Caucasian	15	11	12	11
	African American	0	2	0	1
	Asian	1	0	0	0
Pathological Stage	I-II	\	0	0	0
	III-IV	\	13	12	12
Histology	Adenocarcinoma	\	11	9	10
	Others	\	2	3	2

Table 4. Annotation of the most discriminating cancer type specific DMCGIs

CpG islands	Size (bp)	Gene	Location
Colorectal cancer			
chr2:29337984-29338909	926	*CLIP4*	Promoter
chr2:100937780-100939059	1280	*LONRF2*	Promoter
chr6:125283125-125284389	1265	*RNF217*	Promoter
Lung cancer			
chr2:66672432-66673636	1205	*MEIS1*	Gene body
chr2:71503548-71504233	686	*ZNF638*	Promoter
chr2:219736133-219736592	460	*WNT6*	Gene body
chr4:140655963-140657135	1173	*MGST2*	Gene body
chr4:174427892-174428192	301	\	Intergenic
chr5:40679503-40682081	2579	*PTGER4*	Promoter
chr7:27265159-27265493	335	\	Intergenic
chr7:65037625-65037864	240	\	Intergenic
chr8:124172801-124173541	741	\	Intergenic
chr9:96108467-96108992	526	*C9orf129*	Promoter
chr12:54408427-54408713	287	\	Intergenic
chr12:58021295-58022037	743	*B4GALNT1*	Gene body
chr13:28549840-28550246	407	\	Intergenic
chr17:46691521-46692097	577	*HOXB8*	Gene body
chr17:59539363-59539834	472	*TBX4*	Gene body
chr17:70112825-70114271	1447	*SOX9*	Promoter
Pancreatic cancer			
chr1:44883137-44884272	1136	*RNF220*	Gene body
chr1:50798668-50799536	869	\	Intergenic
chr5:92939796-92940216	421	\	Intergenic
chr10:11059443-11060524	1082	*CELF2*	Promoter
chr11:20177609-20178824	1216	*DBX1*	Gene body
chr12:114881650-114881937	288	\	Intergenic

References

1. Huang, J.; Wang, L. Cell-Free DNA Methylation Profiling Analysis-Technologies and Bioinformatics. *Cancers (Basel)* **2019**, *11*, doi:10.3390/cancers11111741.

2. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer Statistics, 2021. *CA Cancer J Clin* **2021**, *71*, 7-33, doi:10.3322/caac.21654.

3. Force, U.S.P.S.T.; Krist, A.H.; Davidson, K.W.; Mangione, C.M.; Barry, M.J.; Cabana, M.; Caughey, A.B.; Davis, E.M.; Donahue, K.E.; Doubeni, C.A.; et al. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2021**, *325*, 962-970, doi:10.1001/jama.2021.1117.

4. Siu, A.L.; Force, U.S.P.S.T. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med* **2016**, *164*, 279-296, doi:10.7326/M15-2886.

5. Force, U.S.P.S.T.; Curry, S.J.; Krist, A.H.; Owens, D.K.; Barry, M.J.; Caughey, A.B.; Davidson, K.W.; Doubeni, C.A.; Epling, J.W., Jr.; Kemper, A.R.; et al. Screening for Cervical Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2018**, *320*, 674-686, doi:10.1001/jama.2018.10897.

6. Force, U.S.P.S.T.; Davidson, K.W.; Barry, M.J.; Mangione, C.M.; Cabana, M.; Caughey, A.B.; Davis, E.M.; Donahue, K.E.; Doubeni, C.A.; Krist, A.H.; et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2021**, *325*, 1965-1977, doi:10.1001/jama.2021.6238.

7. Force, U.S.P.S.T.; Grossman, D.C.; Curry, S.J.; Owens, D.K.; Bibbins-Domingo, K.; Caughey, A.B.; Davidson, K.W.; Doubeni, C.A.; Ebell, M.; Epling, J.W., Jr.; et al. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2018**, *319*, 1901-1913, doi:10.1001/jama.2018.3710.

8. Pinsky, P.F.; Prorok, P.C.; Kramer, B.S. Prostate Cancer Screening - A Perspective on the Current State of the Evidence. *N Engl J Med* **2017**, *376*, 1285-1289, doi:10.1056/NEJMsb1616281.

9. Wan, J.C.M.; Massie, C.; Garcia-Corbacho, J.; Mouliere, F.; Brenton, J.D.; Caldas, C.; Pacey, S.; Baird, R.; Rosenfeld, N. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* **2017**, *17*, 223-238, doi:10.1038/nrc.2017.7.

10. Siravegna, G.; Marsoni, S.; Siena, S.; Bardelli, A. Integrating liquid biopsies into the management of cancer. *Nature Reviews Clinical Oncology* **2017**, *14*, 531-548, doi:10.1038/nrclinonc.2017.14.

11. Heitzer, E.; Haque, I.S.; Roberts, C.E.S.; Speicher, M.R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat Rev Genet* **2019**, *20*, 71-88, doi:10.1038/s41576-018-0071-5.

12. Jamal-Hanjani, M.; Wilson, G.A.; Horswell, S.; Mitter, R.; Sakarya, O.; Constantin, T.; Salari, R.; Kirkizlar, E.; Sigurjonsson, S.; Pelham, R.; et al. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. *Ann Oncol* **2016**, *27*, 862-867, doi:10.1093/annonc/mdw037.

13. Mattos-Arruda, L.; Weigelt, B.; Cortes, J.; Won, H.H.; Ng, C.K.Y.; Nuciforo, P.; Bidard, F.C.; Aura, C.; Saura, C.; Peg, V.; et al. Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle. *Ann Oncol* **2018**, *29*, 2268, doi:10.1093/annonc/mdx804.

14. Chicard, M.; Boyault, S.; Colmet Daage, L.; Richer, W.; Gentien, D.; Pierron, G.; Lapouble, E.; Bellini, A.; Clement, N.; Iacono, I.; et al. Genomic Copy Number Profiling Using Circulating Free Tumor DNA Highlights Heterogeneity in Neuroblastoma. *Clin Cancer Res* **2016**, *22*, 5564-5573, doi:10.1158/1078-0432.CCR-16-0500.

15. Amorim, M.G.; Valieris, R.; Drummond, R.D.; Pizzi, M.P.; Freitas, V.M.; Sinigaglia-Coimbra, R.; Calin, G.A.; Pasqualini, R.; Arap, W.; Silva, I.T.; et al. A total transcriptome profiling method for plasma-derived extracellular vesicles: applications for liquid biopsies. *Sci Rep* **2017**, *7*, 14395, doi:10.1038/s41598-017-14264-5.

16. Luo, H.; Wei, W.; Ye, Z.; Zheng, J.; Xu, R.H. Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. *Trends Mol Med* **2021**, *27*, 482-500, doi:10.1016/j.molmed.2020.12.011.

17. Gai, W.; Sun, K. Epigenetic Biomarkers in Cell-Free DNA and Applications in Liquid Biopsy. *Genes (Basel)* **2019**, *10*, doi:10.3390/genes10010032.

18. Kim, Y.; Jeon, J.; Mejia, S.; Yao, C.Q.; Ignatchenko, V.; Nyalwidhe, J.O.; Gramolini, A.O.; Lance, R.S.; Troyer, D.A.; Drake, R.R.; et al. Targeted proteomics identifies liquid-biopsy signatures for extracapsular prostate cancer. *Nat Commun* **2016**, *7*, 11906, doi:10.1038/ncomms11906.

19. Mayers, J.R.; Wu, C.; Clish, C.B.; Kraft, P.; Torrence, M.E.; Fiske, B.P.; Yuan, C.; Bao, Y.; Townsend, M.K.; Tworoger, S.S.; et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nat Med* **2014**, *20*, 1193-1198, doi:10.1038/nm.3686.

20. Ko, J.; Baldassano, S.N.; Loh, P.L.; Kording, K.; Litt, B.; Issadore, D. Machine learning to detect signatures of disease in liquid biopsies - a user's guide. *Lab Chip* **2018**, *18*, 395-405, doi:10.1039/c7lc00955k.

21. Chan, K.C.; Jiang, P.; Zheng, Y.W.; Liao, G.J.; Sun, H.; Wong, J.; Siu, S.S.; Chan, W.C.; Chan, S.L.; Chan, A.T.; et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem* **2013**, *59*, 211-224, doi:10.1373/clinchem.2012.196014.

22. Crowley, E.; Di Nicolantonio, F.; Loupakis, F.; Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol* **2013**, *10*, 472-484, doi:10.1038/nrclinonc.2013.110.

23. Thierry, A.R.; El Messaoudi, S.; Gahan, P.B.; Anker, P.; Stroun, M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev* **2016**, *35*, 347-376, doi:10.1007/s10555-016-9629-x.

24. Chandrananda, D.; Thorne, N.P.; Bahlo, M. High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Med Genomics* **2015**, *8*, 29, doi:10.1186/s12920-015-0107-z.

25. Lo, Y.M.; Chan, K.C.; Sun, H.; Chen, E.Z.; Jiang, P.; Lun, F.M.; Zheng, Y.W.; Leung, T.Y.; Lau, T.K.; Cantor, C.R.; et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* **2010**, *2*, 61ra91, doi:10.1126/scitranslmed.3001720.

26. Schwarzenbach, H.; Hoon, D.S.; Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* **2011**, *11*, 426-437, doi:10.1038/nrc3066.

27. Catarino, R.; Ferreira, M.M.; Rodrigues, H.; Coelho, A.; Nogal, A.; Sousa, A.; Medeiros, R. Quantification of free circulating tumor DNA as a diagnostic marker for breast cancer. *DNA Cell Biol* **2008**, *27*, 415-421, doi:10.1089/dna.2008.0744.

28. Bettegowda, C.; Sausen, M.; Leary, R.J.; Kinde, I.; Wang, Y.; Agrawal, N.; Bartlett, B.R.; Wang, H.; Luber, B.; Alani, R.M.; et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **2014**, *6*, 224ra224, doi:10.1126/scitranslmed.3007094.

29. Nygaard, A.D.; Garm Spindler, K.L.; Pallisgaard, N.; Andersen, R.F.; Jakobsen, A. The prognostic value of KRAS mutated plasma DNA in advanced non-small cell lung cancer. *Lung Cancer* **2013**, *79*, 312-317, doi:10.1016/j.lungcan.2012.11.016.

30. Lecomte, T.; Berger, A.; Zinzindohoue, F.; Micard, S.; Landi, B.; Blons, H.; Beaune, P.; Cugnenc, P.H.; Laurent-Puig, P. Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis. *Int J Cancer* **2002**, *100*, 542-548, doi:10.1002/ijc.10526.

31. Gray, E.S.; Rizos, H.; Reid, A.L.; Boyd, S.C.; Pereira, M.R.; Lo, J.; Tembe, V.; Freeman, J.; Lee, J.H.; Scolyer, R.A.; et al. Circulating tumor DNA to monitor treatment response and detect acquired resistance in patients with metastatic melanoma. *Oncotarget* **2015**, *6*, 42008-42018, doi:10.18632/oncotarget.5788.

32. Dawson, S.J.; Tsui, D.W.; Murtaza, M.; Biggs, H.; Rueda, O.M.; Chin, S.F.; Dunning, M.J.; Gale, D.; Forshew, T.; Mahler-Araujo, B.; et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* **2013**, *368*, 1199-1209, doi:10.1056/NEJMoa1213261.

33. Gormally, E.; Vineis, P.; Matullo, G.; Veglia, F.; Caboux, E.; Le Roux, E.; Peluso, M.; Garte, S.; Guarrera, S.; Munnia, A.; et al. TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. *Cancer Res* **2006**, *66*, 6871-6876, doi:10.1158/0008-5472.CAN-05-4556.

34. El Messaoudi, S.; Rolet, F.; Mouliere, F.; Thierry, A.R. Circulating cell free DNA: Preanalytical considerations. *Clin Chim Acta* **2013**, *424*, 222-230, doi:10.1016/j.cca.2013.05.022.

35. Diaz, L.A., Jr.; Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol* **2014**, *32*, 579-586, doi:10.1200/JCO.2012.45.2011.

36. Rolfo, C.; Russo, A. Liquid biopsy for early stage lung cancer moves ever closer. *Nat Rev Clin Oncol* **2020**, *17*, 523-524, doi:10.1038/s41571-020-0393-z.

37. Hu, Y.; Ulrich, B.C.; Supplee, J.; Kuang, Y.; Lizotte, P.H.; Feeney, N.B.; Guibert, N.M.; Awad, M.M.; Wong, K.K.; Janne, P.A.; et al. False-Positive Plasma Genotyping Due to Clonal Hematopoiesis. *Clin Cancer Res* **2018**, *24*, 4437-4443, doi:10.1158/1078-0432.CCR-18-0143.

38. Leary, R.J.; Sausen, M.; Kinde, I.; Papadopoulos, N.; Carpten, J.D.; Craig, D.; O'Shaughnessy, J.; Kinzler, K.W.; Parmigiani, G.; Vogelstein, B.; et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* **2012**, *4*, 162ra154, doi:10.1126/scitranslmed.3004742.

39. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144*, 646-674, doi:10.1016/j.cell.2011.02.013.

40. Dupont, C.; Armant, D.R.; Brenner, C.A. Epigenetics: definition, mechanisms and clinical perspective. *Semin Reprod Med* **2009**, *27*, 351-357, doi:10.1055/s-0029-1237423.

41. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **2012**, *13*, 484-492, doi:10.1038/nrg3230.

42. Ball, M.P.; Li, J.B.; Gao, Y.; Lee, J.H.; LeProust, E.M.; Park, I.H.; Xie, B.; Daley, G.Q.; Church, G.M. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **2009**, *27*, 361-368, doi:10.1038/nbt.1533.

43. Smith, Z.D.; Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **2013**, *14*, 204-220, doi:10.1038/nrg3354.

44. Suzuki, M.M.; Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **2008**, *9*, 465-476, doi:10.1038/nrg2341.

45. Esteller, M. Epigenetics in cancer. *The New England journal of medicine* **2008**, *358*, 1148-1159, doi:10.1056/NEJMra072067.

46. Baylin, S.B.; Jones, P.A. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* **2011**, *11*, 726-734, doi:10.1038/nrc3130.

47. Karpf, A.R.; Matsui, S. Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells. *Cancer Res* **2005**, *65*, 8635-8639, doi:10.1158/0008-5472.CAN-05-1961.

48. Herman, J.G.; Baylin, S.B. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **2003**, *349*, 2042-2054, doi:10.1056/NEJMra023075.

49. Su, J.; Huang, Y.H.; Cui, X.; Wang, X.; Zhang, X.; Lei, Y.; Xu, J.; Lin, X.; Chen, K.; Lv, J.; et al. Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol* **2018**, *19*, 108, doi:10.1186/s13059-018-1492-3.

50. Tao, Y.; Kang, B.; Petkovich, D.A.; Bhandari, Y.R.; In, J.; Stein-O'Brien, G.; Kong, X.; Xie, W.; Zachos, N.; Maegawa, S.; et al. Aging-like Spontaneous Epigenetic Silencing Facilitates Wnt Activation, Stemness, and Braf(V600E)-Induced Tumorigenesis. *Cancer Cell* **2019**, *35*, 315-328 e316, doi:10.1016/j.ccell.2019.01.005.

51. Koch, A.; Joosten, S.C.; Feng, Z.; de Ruijter, T.C.; Draht, M.X.; Melotte, V.; Smits, K.M.; Veeck, J.; Herman, J.G.; Van Neste, L.; et al. Analysis of DNA methylation in cancer: location revisited. *Nat Rev Clin Oncol* **2018**, *15*, 459-466, doi:10.1038/s41571-018-0004-4.

52. Laird, P.W. The power and the promise of DNA methylation markers. *Nat Rev Cancer* **2003**, *3*, 253-266, doi:10.1038/nrc1045.

53. Takeshima, H.; Ushijima, T. Accumulation of genetic and epigenetic alterations in normal cells and cancer risk. *NPJ Precis Oncol* **2019**, *3*, 7, doi:10.1038/s41698-019-0079-0.

54. Widschwendter, M.; Jones, A.; Evans, I.; Reisel, D.; Dillner, J.; Sundstrom, K.; Steyerberg, E.W.; Vergouwe, Y.; Wegwarth, O.; Rebitschek, F.G.; et al. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nat Rev Clin Oncol* **2018**, *15*, 292-309, doi:10.1038/nrclinonc.2018.30.

55. Robertson, K.D. DNA methylation, methyltransferases, and cancer. *Oncogene* **2001**, *20*, 3139-3155, doi:10.1038/sj.onc.1204341.

56. Roadmap Epigenomics, C.; Kundaje, A.; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; et al. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317-330, doi:10.1038/nature14248.

57. Moss, J.; Magenheim, J.; Neiman, D.; Zemmour, H.; Loyfer, N.; Korach, A.; Samet, Y.; Maoz, M.; Druid, H.; Arner, P.; et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* **2018**, *9*, 5068, doi:10.1038/s41467-018-07466-6.

58. Board, R.E.; Knight, L.; Greystoke, A.; Blackhall, F.H.; Hughes, A.; Dive, C.; Ranson, M. DNA methylation in circulating tumour DNA as a biomarker for cancer. *Biomark Insights* **2008**, *2*, 307-319.

59. Feng, H.; Jin, P.; Wu, H. Disease prediction by cell-free DNA methylation. *Brief Bioinform* **2019**, *20*, 585-597, doi:10.1093/bib/bby029.

60. Locke, W.J.; Guanzon, D.; Ma, C.; Liew, Y.J.; Duesing, K.R.; Fung, K.Y.C.; Ross, J.P. DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Front Genet* **2019**, *10*, 1150, doi:10.3389/fgene.2019.01150.

61. Dam-Dcm and CpG Methylation. Available online: https://www.neb.com/tools-and-resources/selection-charts/dam-dcm-and-cpg-methylation (accessed on).

62. Oda, M.; Glass, J.L.; Thompson, R.F.; Mo, Y.; Olivier, E.N.; Figueroa, M.E.; Selzer, R.R.; Richmond, T.A.; Zhang, X.; Dannenberg, L.; et al. High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res* **2009**, *37*, 3829-3839, doi:10.1093/nar/gkp260.

63. Irizarry, R.A.; Ladd-Acosta, C.; Carvalho, B.; Wu, H.; Brandenburg, S.A.; Jeddeloh, J.A.; Wen, B.; Feinberg, A.P. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* **2008**, *18*, 780-790, doi:10.1101/gr.7301508.

64. Maunakea, A.K.; Nagarajan, R.P.; Bilenky, M.; Ballinger, T.J.; D'Souza, C.; Fouse, S.D.; Johnson, B.E.; Hong, C.; Nielsen, C.; Zhao, Y.; et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **2010**, *466*, 253-257, doi:10.1038/nature09165.

65. Brunner, A.L.; Johnson, D.S.; Kim, S.W.; Valouev, A.; Reddy, T.E.; Neff, N.F.; Anton, E.; Medina, C.; Nguyen, L.; Chiao, E.; et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res* **2009**, *19*, 1044-1056, doi:10.1101/gr.088773.108.

66. Boers, R.; Boers, J.; de Hoon, B.; Kockx, C.; Ozgur, Z.; Molijn, A.; van, I.W.; Laven, J.; Gribnau, J. Genome-wide DNA methylation profiling using the methylation-dependent restriction enzyme LpnPI. *Genome Res* **2018**, *28*, 88-99, doi:10.1101/gr.222885.117.

67. Deger, T.; Boers, R.G.; de Weerd, V.; Angus, L.; van der Put, M.M.J.; Boers, J.B.; Azmani, Z.; van, I.W.F.J.; Grunhagen, D.J.; van Dessel, L.F.; et al. High-throughput and affordable genome-wide methylation profiling of circulating cell-free DNA by methylated DNA sequencing (MeD-seq) of LpnPI digested fragments. *Clin Epigenetics* **2021**, *13*, 196, doi:10.1186/s13148-021-01177-4.

68. Frommer, M.; McDonald, L.E.; Millar, D.S.; Collis, C.M.; Watt, F.; Grigg, G.W.; Molloy, P.L.; Paul, C.L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **1992**, *89*, 1827-1831, doi:10.1073/pnas.89.5.1827.

69. Tanaka, K.; Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg Med Chem Lett* **2007**, *17*, 1912-1915, doi:10.1016/j.bmcl.2007.01.040.

70. Beck, S.; Rakyan, V.K. The methylome: approaches for global DNA methylation profiling. *Trends in Genetics* **2008**, *24*, 231-237, doi:10.1016/j.tig.2008.01.006.

71. Lister, R.; Pelizzola, M.; Dowen, R.H.; Hawkins, R.D.; Hon, G.; Tonti-Filippini, J.; Nery, J.R.; Lee, L.; Ye, Z.; Ngo, Q.M.; et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **2009**, *462*, 315-322, doi:10.1038/nature08514.

72. Smallwood, S.A.; Lee, H.J.; Angermueller, C.; Krueger, F.; Saadeh, H.; Peat, J.; Andrews, S.R.; Stegle, O.; Reik, W.; Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **2014**, *11*, 817-820, doi:10.1038/nmeth.3035.

73. Farlik, M.; Sheffield, N.C.; Nuzzo, A.; Datlinger, P.; Schonegger, A.; Klughammer, J.; Bock, C. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* **2015**, *10*, 1386-1397, doi:10.1016/j.celrep.2015.02.001.

74. Miura, F.; Enomoto, Y.; Dairiki, R.; Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* **2012**, *40*, e136, doi:10.1093/nar/gks454.

75. Miura, F.; Shibata, Y.; Miura, M.; Sangatsuda, Y.; Hisano, O.; Araki, H.; Ito, T. Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* **2019**, *47*, e85, doi:10.1093/nar/gkz435.

76. Clark, S.J.; Smallwood, S.A.; Lee, H.J.; Krueger, F.; Reik, W.; Kelsey, G. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc* **2017**, *12*, 534-547, doi:10.1038/nprot.2016.187.

77. Legendre, C.; Gooden, G.C.; Johnson, K.; Martinez, R.A.; Liang, W.S.; Salhia, B. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clin Epigenetics* **2015**, *7*, 100, doi:10.1186/s13148-015-0135-8.

78. Liu, M.C.; Oxnard, G.R.; Klein, E.A.; Swanton, C.; Seiden, M.V.; Consortium, C. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol* **2020**, *31*, 745-759, doi:10.1016/j.annonc.2020.02.011.

79. Meissner, A.; Mikkelsen, T.S.; Gu, H.; Wernig, M.; Hanna, J.; Sivachenko, A.; Zhang, X.; Bernstein, B.E.; Nusbaum, C.; Jaffe, D.B.; et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **2008**, *454*, 766-770, doi:10.1038/nature07107.

80. Gu, H.; Smith, Z.D.; Bock, C.; Boyle, P.; Gnirke, A.; Meissner, A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **2011**, *6*, 468-481, doi:10.1038/nprot.2010.190.

81. Smith, Z.D.; Gu, H.; Bock, C.; Gnirke, A.; Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **2009**, *48*, 226-232, doi:10.1016/j.ymeth.2009.05.003.

82. Guo, H.; Zhu, P.; Wu, X.; Li, X.; Wen, L.; Tang, F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **2013**, *23*, 2126-2135, doi:10.1101/gr.161679.113.

83. Guo, H.; Zhu, P.; Guo, F.; Li, X.; Wu, X.; Fan, X.; Wen, L.; Tang, F. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat Protoc* **2015**, *10*, 645-659, doi:10.1038/nprot.2015.039.

84. Guo, S.; Diep, D.; Plongthongkum, N.; Fung, H.L.; 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*, 635-642, doi:10.1038/ng.3805.

85. Van Paemel R., D.K.A., Vandeputte C., van Zogchel L., Lammens T., Laureys G., et al. Minimally invasive classification of paediatric solid tumours using reduced representation bisulphite sequencing of cell-free DNA: a proof-of-principle study. *Epigenetics 1–13. 10.1080/15592294.2020.1790950.* **2020**, doi:https://doi.org/10.1080/15592294.2020.1790950.

86. Wen, L.; Li, J.; Guo, H.; Liu, X.; Zheng, S.; Zhang, D.; Zhu, W.; Qu, J.; Guo, L.; Du, D.; et al. Genome-scale detection of hypermethylated CpG islands in circulating cell-free DNA of hepatocellular carcinoma patients. *Cell Res* **2015**, *25*, 1376, doi:10.1038/cr.2015.141.

87. Li, J.; Zhou, X.; Liu, X.; Ren, J.; Wang, J.; Wang, W.; Zheng, Y.; Shi, X.; Sun, T.; Li, Z.; et al. Detection of Colorectal Cancer in Circulating Cell-Free DNA by Methylated CpG Tandem Amplification and Sequencing. *Clin Chem* **2019**, doi:10.1373/clinchem.2019.301804.

88. Ren, J.; Lu, P.; Zhou, X.; Liao, Y.; Liu, X.; Li, J.; Wang, W.; Wang, J.; Wen, L.; Fu, W.; et al. Genome-Scale Methylation Analysis of Circulating Cell-Free DNA in Gastric Cancer Patients. *Clin Chem* **2021**, doi:10.1093/clinchem/hvab204.

89. Liu, X.; Ren, J.; Luo, N.; Guo, H.; Zheng, Y.; Li, J.; Tang, F.; Wen, L.; Peng, J. Comprehensive DNA methylation analysis of tissue of origin of plasma cell-free DNA by

methylated CpG tandem amplification and sequencing (MCTA-Seq). *Clin Epigenetics* **2019**, *11*, 93, doi:10.1186/s13148-019-0689-y.

90. Samorodnitsky, E.; Datta, J.; Jewell, B.M.; Hagopian, R.; Miya, J.; Wing, M.R.; Damodaran, S.; Lippus, J.M.; Reeser, J.W.; Bhatt, D.; et al. Comparison of custom capture for targeted next-generation DNA sequencing. *J Mol Diagn* **2015**, *17*, 64-75, doi:10.1016/j.jmoldx.2014.09.009.

91. Widschwendter, M.; Evans, I.; Jones, A.; Ghazali, S.; Reisel, D.; Ryan, A.; Gentry-Maharaj, A.; Zikan, M.; Cibula, D.; Eichner, J.; et al. Methylation patterns in serum DNA for early identification of disseminated breast cancer. *Genome Med* **2017**, *9*, 115, doi:10.1186/s13073-017-0499-9.

92. Holmila, R.; Sklias, A.; Muller, D.C.; Degli Esposti, D.; Guilloreau, P.; McKay, J.; Sangrajrang, S.; Srivatanakul, P.; Hainaut, P.; Merle, P.; et al. Targeted deep sequencing of plasma circulating cell-free DNA reveals Vimentin and Fibulin 1 as potential epigenetic biomarkers for hepatocellular carcinoma. *PLoS One* **2017**, *12*, e0174265, doi:10.1371/journal.pone.0174265.

93. Liu, L.; Toung, J.M.; Jassowicz, A.F.; Vijayaraghavan, R.; Kang, H.; Zhang, R.; Kruglyak, K.M.; Huang, H.J.; Hinoue, T.; Shen, H.; et al. Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification. *Ann Oncol* **2018**, *29*, 1445-1453, doi:10.1093/annonc/mdy119.

94. Bibikova, M.; Barnes, B.; Tsan, C.; Ho, V.; Klotzle, B.; Le, J.M.; Delano, D.; Zhang, L.; Schroth, G.P.; Gunderson, K.L.; et al. High density DNA methylation array with single CpG site resolution. *Genomics* **2011**, *98*, 288-295, doi:10.1016/j.ygeno.2011.07.007.

95. Stirzaker, C.; Taberlay, P.C.; Statham, A.L.; Clark, S.J. Mining cancer methylomes: prospects and challenges. *Trends Genet* **2014**, *30*, 75-84, doi:10.1016/j.tig.2013.11.004.

96. Moran, S.; Arribas, C.; Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **2016**, *8*, 389-399, doi:10.2217/epi.15.114.

97. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **2013**, *41*, D991-995, doi:10.1093/nar/gks1193.

98. The Cancer Genome Atlas. Available online: https://www.cancer.gov/tcga (accessed on).

99. Vrba, L.; Futscher, B.W. A suite of DNA methylation markers that can detect most common human cancers. *Epigenetics* **2018**, *13*, 61-72, doi:10.1080/15592294.2017.1412907.

100. Hao, X.K.; Luo, H.Y.; Krawczyk, M.; Wei, W.; Wang, W.Q.; Wang, J.; Flagg, K.; Hou, J.Y.; Zhang, H.; Yi, S.H.; et al. DNA methylation markers for diagnosis and prognosis of common cancers. *P Natl Acad Sci USA* **2017**, *114*, 7414-7419, doi:10.1073/pnas.1703577114.

101. Huang, J.; Soupir, A.C.; Wang, L. Cell-free DNA methylome profiling by MBD-seq with ultra-low input. *Epigenetics* **2021**, 1-14, doi:10.1080/15592294.2021.1896984.

102. Gallardo-Gomez, M.; Moran, S.; Paez de la Cadena, M.; Martinez-Zorzano, V.S.; Rodriguez-Berrocal, F.J.; Rodriguez-Girondo, M.; Esteller, M.; Cubiella, J.; Bujanda, L.; Castells, A.; et al. A new approach to epigenome-wide discovery of non-invasive methylation biomarkers for colorectal cancer screening in circulating cell-free DNA using pooled samples. *Clin Epigenetics* **2018**, *10*, 53, doi:10.1186/s13148-018-0487-y.

103. Hlady, R.A.; Zhao, X.; Pan, X.; Yang, J.D.; Ahmed, F.; Antwi, S.O.; Giama, N.H.; Patel, T.; Roberts, L.R.; Liu, C.; et al. Genome-wide discovery and validation of diagnostic DNA

methylation-based biomarkers for hepatocellular cancer detection in circulating cell free DNA. *Theranostics* **2019**, *9*, 7239-7250, doi:10.7150/thno.35573.

104. Gordevicius, J.; Krisciunas, A.; Groot, D.E.; Yip, S.M.; Susic, M.; Kwan, A.; Kustra, R.; Joshua, A.M.; Chi, K.N.; Petronis, A.; et al. Cell-Free DNA Modification Dynamics in Abiraterone Acetate-Treated Prostate Cancer Patients. *Clin Cancer Res* **2018**, *24*, 3317-3324, doi:10.1158/1078-0432.CCR-18-0101.

105. Herman, J.G.; Graff, J.R.; Myohanen, S.; Nelkin, B.D.; Baylin, S.B. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A* **1996**, *93*, 9821-9826, doi:10.1073/pnas.93.18.9821.

106. Eads, C.A.; Danenberg, K.D.; Kawakami, K.; Saltz, L.B.; Blake, C.; Shibata, D.; Danenberg, P.V.; Laird, P.W. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res* **2000**, *28*, E32, doi:10.1093/nar/28.8.e32.

107. Lo, P.K.; Watanabe, H.; Cheng, P.C.; Teo, W.W.; Liang, X.; Argani, P.; Lee, J.S.; Sukumar, S. MethySYBR, a novel quantitative PCR assay for the dual analysis of DNA methylation and CpG methylation density. *J Mol Diagn* **2009**, *11*, 400-414, doi:10.2353/jmoldx.2009.080126.

108. Dugast-Darzacq, C.; Grange, T. MethylQuant: a real-time PCR-based method to quantify DNA methylation at single specific cytosines. *Methods Mol Biol* **2009**, *507*, 281-303, doi:10.1007/978-1-59745-522-0_21.

109. Wojdacz, T.K.; Dobrovic, A. Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. *Nucleic Acids Res* **2007**, *35*, e41, doi:10.1093/nar/gkm013.

110. Panagopoulou, M.; Karaglani, M.; Balgkouranidou, I.; Biziota, E.; Koukaki, T.; Karamitrousis, E.; Nena, E.; Tsamardinos, I.; Kolios, G.; Lianidou, E.; et al. Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. *Oncogene* **2019**, *38*, 3387-3401, doi:10.1038/s41388-018-0660-y.

111. Kloten, V.; Becker, B.; Winner, K.; Schrauder, M.G.; Fasching, P.A.; Anzeneder, T.; Veeck, J.; Hartmann, A.; Knuchel, R.; Dahl, E. Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based breast cancer screening. *Breast Cancer Res* **2013**, *15*, R4, doi:10.1186/bcr3375.

112. Hung, C.S.; Wang, S.C.; Yen, Y.T.; Lee, T.H.; Wen, W.C.; Lin, R.K. Hypermethylation of CCND2 in Lung and Breast Cancer Is a Potential Biomarker and Drug Target. *Int J Mol Sci* **2018**, *19*, doi:10.3390/ijms19103096.

113. Eissa, M.A.L.; Lerner, L.; Abdelfatah, E.; Shankar, N.; Canner, J.K.; Hasan, N.M.; Yaghoobi, V.; Huang, B.; Kerner, Z.; Takaesu, F.; et al. Promoter methylation of ADAMTS1 and BNC1 as potential biomarkers for early detection of pancreatic cancer in blood. *Clin Epigenetics* **2019**, *11*, 59, doi:10.1186/s13148-019-0650-0.

114. Giannopoulou, L.; Mastoraki, S.; Buderath, P.; Strati, A.; Pavlakis, K.; Kasimir-Bauer, S.; Lianidou, E.S. ESR1 methylation in primary tumors and paired circulating tumor DNA of patients with high-grade serous ovarian cancer. *Gynecol Oncol* **2018**, *150*, 355-360, doi:10.1016/j.ygyno.2018.05.026.

115. Wasenang, W.; Chaiyarit, P.; Proungvitaya, S.; Limpaiboon, T. Serum cell-free DNA methylation of OPCML and HOXD9 as a biomarker that may aid in differential diagnosis between cholangiocarcinoma and other biliary diseases. *Clin Epigenetics* **2019**, *11*, 39, doi:10.1186/s13148-019-0634-0.

116. Zhao, Y.; Xue, F.; Sun, J.; Guo, S.; Zhang, H.; Qiu, B.; Geng, J.; Gu, J.; Zhou, X.; Wang, W.; 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*, 30, doi:10.1186/1868-7083-6-30.

117. Oussalah, A.; Rischer, S.; Bensenane, M.; Conroy, G.; Filhine-Tresarrieu, P.; Debard, R.; Forest-Tramoy, D.; Josse, T.; Reinicke, D.; Garcia, M.; et al. Plasma mSEPT9: A Novel Circulating Cell-free DNA-Based Epigenetic Biomarker to Diagnose Hepatocellular Carcinoma. *EBioMedicine* **2018**, *30*, 138-147, doi:10.1016/j.ebiom.2018.03.029.

118. Bjerre, M.T.; Strand, S.H.; Norgaard, M.; Kristensen, H.; Rasmussen, A.K.; Mortensen, M.M.; Fredsoe, J.; Mouritzen, P.; Ulhoi, B.; Orntoft, T.; et al. Aberrant DOCK2, GRASP, HIF3A and PKFP Hypermethylation has Potential as a Prognostic Biomarker for Prostate Cancer. *Int J Mol Sci* **2019**, *20*, doi:10.3390/ijms20051173.

119. Powrozek, T.; Krawczyk, P.; Nicos, M.; Kuznar-Kaminska, B.; Batura-Gabryel, H.; Milanowski, J. Methylation of the DCLK1 promoter region in circulating free DNA and its prognostic value in lung cancer patients. *Clin Transl Oncol* **2016**, *18*, 398-404, doi:10.1007/s12094-015-1382-z.

120. Nunes, S.P.; Moreira-Barbosa, C.; Salta, S.; Palma de Sousa, S.; Pousa, I.; Oliveira, J.; Soares, M.; Rego, L.; Dias, T.; Rodrigues, J.; et al. Cell-Free DNA Methylation of Selected Genes Allows for Early Detection of the Major Cancers in Women. *Cancers (Basel)* **2018**, *10*, doi:10.3390/cancers10100357.

121. Uehiro, N.; Sato, F.; Pu, F.; Tanaka, S.; Kawashima, M.; Kawaguchi, K.; Sugimoto, M.; Saji, S.; Toi, M. Circulating cell-free DNA-based epigenetic assay can detect early breast cancer. *Breast Cancer Res* **2016**, *18*, 129, doi:10.1186/s13058-016-0788-z.

122. Haldrup, C.; Pedersen, A.L.; Ogaard, N.; Strand, S.H.; Hoyer, S.; Borre, M.; Orntoft, T.F.; Sorensen, K.D. Biomarker potential of ST6GALNAC3 and ZNF660 promoter hypermethylation in prostate cancer tissue and liquid biopsies. *Mol Oncol* **2018**, *12*, 545-560, doi:10.1002/1878-0261.12183.

123. Picardo, F.; Romanelli, A.; Muinelo-Romay, L.; Mazza, T.; Fusilli, C.; Parrella, P.; Barbazan, J.; Lopez-Lopez, R.; Barbano, R.; De Robertis, M.; et al. Diagnostic and Prognostic Value of B4GALT1 Hypermethylation and Its Clinical Significance as a Novel Circulating Cell-Free DNA Biomarker in Colorectal Cancer. *Cancers (Basel)* **2019**, *11*, doi:10.3390/cancers11101598.

124. Barault, L.; Amatu, A.; Bleeker, F.E.; Moutinho, C.; Falcomata, C.; Fiano, V.; Cassingena, A.; Siravegna, G.; Milione, M.; Cassoni, P.; et al. Digital PCR quantification of MGMT methylation refines prediction of clinical benefit from alkylating agents in glioblastoma and metastatic colorectal cancer. *Ann Oncol* **2015**, *26*, 1994-1999, doi:10.1093/annonc/mdv272.

125. Jensen, S.O.; Ogaard, N.; Orntoft, M.W.; Rasmussen, M.H.; Bramsen, J.B.; Kristensen, H.; Mouritzen, P.; Madsen, M.R.; Madsen, A.H.; Sunesen, K.G.; et al. Novel DNA methylation biomarkers show high sensitivity and specificity for blood-based detection of colorectal cancer-a clinical biomarker discovery and validation study. *Clin Epigenetics* **2019**, *11*, 158, doi:10.1186/s13148-019-0757-3.

126. Tahiliani, M.; Koh, K.P.; Shen, Y.; Pastor, W.A.; Bandukwala, H.; Brudno, Y.; Agarwal, S.; Iyer, L.M.; Liu, D.R.; Aravind, L.; et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **2009**, *324*, 930-935, doi:10.1126/science.1170116.

127. He, Y.F.; Li, B.Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, L.; et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **2011**, *333*, 1303-1307, doi:10.1126/science.1210944.

128. Booth, M.J.; Ost, T.W.; Beraldi, D.; Bell, N.M.; Branco, M.R.; Reik, W.; Balasubramanian, S. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* **2013**, *8*, 1841-1851, doi:10.1038/nprot.2013.115.

129. Yu, M.; Hon, G.C.; Szulwach, K.E.; Song, C.X.; Jin, P.; Ren, B.; He, C. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat Protoc* **2012**, *7*, 2159-2170, doi:10.1038/nprot.2012.137.

130. Barros-Silva, D.; Marques, C.J.; Henrique, R.; Jeronimo, C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. *Genes (Basel)* **2018**, *9*, doi:10.3390/genes9090429.

131. Chan, R.F.; Shabalin, A.A.; Xie, L.Y.; Adkins, D.E.; Zhao, M.; Turecki, G.; Clark, S.L.; Aberg, K.A.; van den Oord, E. Enrichment methods provide a feasible approach to comprehensive and adequately powered investigations of the brain methylome. *Nucleic Acids Res* **2017**, *45*, e97, doi:10.1093/nar/gkx143.

132. Weber, M.; Davies, J.J.; Wittig, D.; Oakeley, E.J.; Haase, M.; Lam, W.L.; Schubeler, D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **2005**, *37*, 853-862, doi:10.1038/ng1598.

133. Taiwo, O.; Wilson, G.A.; Morris, T.; Seisenberger, S.; Reik, W.; Pearce, D.; Beck, S.; Butcher, L.M. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* **2012**, *7*, 617-636, doi:10.1038/nprot.2012.012.

134. Shen, S.Y.; Singhania, R.; Fehringer, G.; Chakravarthy, A.; Roehrl, M.H.A.; Chadwick, D.; Zuzarte, P.C.; Borgida, A.; Wang, T.T.; Li, T.T.; et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **2018**, *563*, 579-+, doi:10.1038/s41586-018-0703-0.

135. Shen, S.Y.; Burgener, J.M.; Bratman, S.V.; De Carvalho, D.D. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nat Protoc* **2019**, *14*, 2749-2780, doi:10.1038/s41596-019-0202-2.

136. Nuzzo, P.V.; Berchuck, J.E.; Korthauer, K.; Spisak, S.; Nassar, A.H.; Abou Alaiwi, S.; Chakravarthy, A.; Shen, S.Y.; Bakouny, Z.; Boccardo, F.; et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat Med* **2020**, *26*, 1041-1043, doi:10.1038/s41591-020-0933-1.

137. Lasseter, K.; Nassar, A.H.; Hamieh, L.; Berchuck, J.E.; Nuzzo, P.V.; Korthauer, K.; Shinagare, A.B.; Ogorek, B.; McKay, R.; Thorner, A.R.; et al. Plasma cell-free DNA variant analysis compared with methylated DNA analysis in renal cell carcinoma. *Genet Med* **2020**, doi:10.1038/s41436-020-0801-x.

138. Farshad Nassiri, A.C., Shengrui Feng, Shu Yi Shen, Romina Nejad, Jeffrey A. Zuccato, Mathew R. Voisin, Vikas Patil, Craig Horbinski, Kenneth Aldape, Gelareh Zadeh & Daniel D. De Carvalho. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nature Medicine volume 26, pages 1044–1047 (2020)* **2020**, doi:https://doi.org/10.1038/s41591-020-0932-2.

139. Brinkman, A.B.; Simmer, F.; Ma, K.; Kaan, A.; Zhu, J.; Stunnenberg, H.G. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods* **2010**, *52*, 232-236, doi:10.1016/j.ymeth.2010.06.012.

140. Nair, S.S.; Coolen, M.W.; Stirzaker, C.; Song, J.Z.; Statham, A.L.; Strbenac, D.; Robinson, M.D.; Clark, S.J. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* **2011**, *6*, 34-44, doi:10.4161/epi.6.1.13313.

141. Aberg, K.A.; Chan, R.F.; Shabalin, A.A.; Zhao, M.; Turecki, G.; Staunstrup, N.H.; Starnawska, A.; Mors, O.; Xie, L.Y.; van den Oord, E.J. A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. *Epigenetics* **2017**, *12*, 743-750, doi:10.1080/15592294.2017.1335849.

142. Bachman, M.; Uribe-Lewis, S.; Yang, X.; Williams, M.; Murrell, A.; Balasubramanian, S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* **2014**, *6*, 1049-1055, doi:10.1038/nchem.2064.

143. Vasanthakumar, A.; Godley, L.A. 5-hydroxymethylcytosine in cancer: significance in diagnosis and therapy. *Cancer Genet* **2015**, *208*, 167-177, doi:10.1016/j.cancergen.2015.02.009.

144. Han, D.; Lu, X.; Shih, A.H.; Nie, J.; You, Q.; Xu, M.M.; Melnick, A.M.; Levine, R.L.; He, C. A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Mol Cell* **2016**, *63*, 711-719, doi:10.1016/j.molcel.2016.06.028.

145. Song, C.X.; Szulwach, K.E.; Fu, Y.; Dai, Q.; Yi, C.; Li, X.; Li, Y.; Chen, C.H.; Zhang, W.; Jian, X.; et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **2011**, *29*, 68-72, doi:10.1038/nbt.1732.

146. Song, C.X.; Yin, S.; Ma, L.; Wheeler, A.; Chen, Y.; Zhang, Y.; Liu, B.; Xiong, J.; Zhang, W.; Hu, J.; et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res* **2017**, *27*, 1231-1242, doi:10.1038/cr.2017.106.

147. Li, W.; Zhang, X.; Lu, X.; You, L.; Song, Y.; Luo, Z.; Zhang, J.; Nie, J.; Zheng, W.; Xu, D.; et al. 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res* **2017**, *27*, 1243-1257, doi:10.1038/cr.2017.121.

148. Zhang, J.; Han, X.; Gao, C.; Xing, Y.; Qi, Z.; Liu, R.; Wang, Y.; Zhang, X.; Yang, Y.G.; Li, X.; et al. 5-Hydroxymethylome in Circulating Cell-free DNA as A Potential Biomarker for Non-small-cell Lung Cancer. *Genomics Proteomics Bioinformatics* **2018**, *16*, 187-199, doi:10.1016/j.gpb.2018.06.002.

149. Gao, P.; Lin, S.; Cai, M.; Zhu, Y.; Song, Y.; Sui, Y.; Lin, J.; Liu, J.; Lu, X.; Zhong, Y.; et al. 5-Hydroxymethylcytosine profiling from genomic and cell-free DNA for colorectal cancers patients. *J Cell Mol Med* **2019**, *23*, 3530-3537, doi:10.1111/jcmm.14252.

150. Cai, J.; Chen, L.; Zhang, Z.; Zhang, X.; Lu, X.; Liu, W.; Shi, G.; Ge, Y.; Gao, P.; Yang, Y.; et al. Genome-wide mapping of 5-hydroxymethylcytosines in circulating cell-free DNA as a non-invasive approach for early detection of hepatocellular carcinoma. *Gut* **2019**, doi:10.1136/gutjnl-2019-318882.

151. Tian, X.; Sun, B.; Chen, C.; Gao, C.; Zhang, J.; Lu, X.; Wang, L.; Li, X.; Xing, Y.; Liu, R.; et al. Circulating tumor DNA 5-hydroxymethylcytosine as a novel diagnostic biomarker for esophageal cancer. *Cell Res* **2018**, *28*, 597-600, doi:10.1038/s41422-018-0014-x.

152. FastQC. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on).

153. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114-2120, doi:10.1093/bioinformatics/btu170.

154. Trim Galore. Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on).

155. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884-i890, doi:10.1093/bioinformatics/bty560.
156. Xi, Y.; Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *Bmc Bioinformatics* **2009**, *10*, 232, doi:10.1186/1471-2105-10-232.
157. Xi, Y.; Bock, C.; Muller, F.; Sun, D.; Meissner, A.; Li, W. RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics* **2012**, *28*, 430-432, doi:10.1093/bioinformatics/btr668.
158. Coarfa, C.; Yu, F.; Miller, C.A.; Chen, Z.; Harris, R.A.; Milosavljevic, A. Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. *Bmc Bioinformatics* **2010**, *11*, 572, doi:10.1186/1471-2105-11-572.
159. Wu, T.D.; Reeder, J.; Lawrence, M.; Becker, G.; Brauer, M.J. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods Mol Biol* **2016**, *1418*, 283-334, doi:10.1007/978-1-4939-3578-9_15.
160. Krueger, F.; Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **2011**, *27*, 1571-1572, doi:10.1093/bioinformatics/btr167.
161. Huang, K.Y.Y.; Huang, Y.J.; Chen, P.Y. BS-Seeker3: ultrafast pipeline for bisulfite sequencing. *Bmc Bioinformatics* **2018**, *19*, 111, doi:10.1186/s12859-018-2120-7.
162. Harris, E.Y.; Ponts, N.; Le Roch, K.G.; Lonardi, S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. *Bioinformatics* **2012**, *28*, 1795-1796, doi:10.1093/bioinformatics/bts264.
163. Kunde-Ramamoorthy, G.; Coarfa, C.; Laritsky, E.; Kessler, N.J.; Harris, R.A.; Xu, M.; Chen, R.; Shen, L.; Milosavljevic, A.; Waterland, R.A. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res* **2014**, *42*, e43, doi:10.1093/nar/gkt1325.
164. Bock, C. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics* **2012**, *13*, 705-719, doi:10.1038/nrg3273.
165. Daca-Roszak, P.; Pfeifer, A.; Zebracka-Gala, J.; Rusinek, D.; Szybinska, A.; Jarzab, B.; Witt, M.; Zietkiewicz, E. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies. *BMC Genomics* **2015**, *16*, 1003, doi:10.1186/s12864-015-2202-0.
166. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **2012**, *9*, 357-359, doi:10.1038/nmeth.1923.
167. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754-1760, doi:10.1093/bioinformatics/btp324.
168. Lienhard, M.; Grimm, C.; Morkel, M.; Herwig, R.; Chavez, L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **2014**, *30*, 284-286, doi:10.1093/bioinformatics/btt650.
169. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **2008**, *5*, 621-628, doi:10.1038/nmeth.1226.
170. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **2010**, *11*, R25, doi:10.1186/gb-2010-11-3-r25.
171. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **2010**, *11*, R106, doi:10.1186/gb-2010-11-10-r106.

172. Down, T.A.; Rakyan, V.K.; Turner, D.J.; Flicek, P.; Li, H.; Kulesha, E.; Graf, S.; Johnson, N.; Herrero, J.; Tomazou, E.M.; et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **2008**, *26*, 779-785, doi:10.1038/nbt1414.
173. Lienhard, M.; Grasse, S.; Rolff, J.; Frese, S.; Schirmer, U.; Becker, M.; Borno, S.; Timmermann, B.; Chavez, L.; Sultmann, H.; et al. QSEA-modelling of genome-wide DNA methylation from sequencing enrichment experiments. *Nucleic Acids Res* **2017**, *45*, e44, doi:10.1093/nar/gkw1193.
174. van den Oord, E.J.; Bukszar, J.; Rudolf, G.; Nerella, S.; McClay, J.L.; Xie, L.Y.; Aberg, K.A. Estimation of CpG coverage in whole methylome next-generation sequencing studies. *Bmc Bioinformatics* **2013**, *14*, 50, doi:10.1186/1471-2105-14-50.
175. Shabalin, A.A.; Hattab, M.W.; Clark, S.L.; Chan, R.F.; Kumar, G.; Aberg, K.A.; van den Oord, E. RaMWAS: fast methylome-wide association study pipeline for enrichment platforms. *Bioinformatics* **2018**, *34*, 2283-2285, doi:10.1093/bioinformatics/bty069.
176. Haeussler, M.; Zweig, A.S.; Tyner, C.; Speir, M.L.; Rosenbloom, K.R.; Raney, B.J.; Lee, C.M.; Lee, B.T.; Hinrichs, A.S.; Gonzalez, J.N.; et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **2019**, *47*, D853-D858, doi:10.1093/nar/gky1095.
177. Thorvaldsdottir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **2013**, *14*, 178-192, doi:10.1093/bib/bbs017.
178. Mallona, I.; Diez-Villanueva, A.; Peinado, M.A. Methylation plotter: a web tool for dynamic visualization of DNA methylation data. *Source Code Biol Med* **2014**, *9*, 11, doi:10.1186/1751-0473-9-11.
179. Liang, F.; Tang, B.; Wang, Y.; Wang, J.; Yu, C.; Chen, X.; Zhu, J.; Yan, J.; Zhao, W.; Li, R. WBSA: web service for bisulfite sequencing data analysis. *PLoS One* **2014**, *9*, e86707, doi:10.1371/journal.pone.0086707.
180. Hansen, K.D.; Langmead, B.; Irizarry, R.A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **2012**, *13*, R83, doi:10.1186/gb-2012-13-10-r83.
181. Aryee, M.J.; Jaffe, A.E.; Corrada-Bravo, H.; Ladd-Acosta, C.; Feinberg, A.P.; Hansen, K.D.; Irizarry, R.A. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **2014**, *30*, 1363-1369, doi:10.1093/bioinformatics/btu049.
182. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **2004**, *3*, Article3, doi:10.2202/1544-6115.1027.
183. Xu, H.; Podolsky, R.H.; Ryu, D.; Wang, X.; Su, S.; Shi, H.; George, V. A method to detect differentially methylated loci with next-generation sequencing. *Genet Epidemiol* **2013**, *37*, 377-382, doi:10.1002/gepi.21726.
184. Feng, H.; Conneely, K.N.; Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* **2014**, *42*, e69, doi:10.1093/nar/gku154.
185. Park, Y.; Figueroa, M.E.; Rozek, L.S.; Sartor, M.A. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* **2014**, *30*, 2414-2422, doi:10.1093/bioinformatics/btu339.

186. Dolzhenko, E.; Smith, A.D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *Bmc Bioinformatics* **2014**, *15*, 215, doi:10.1186/1471-2105-15-215.

187. Warden, C.D.; Lee, H.; Tompkins, J.D.; Li, X.; Wang, C.; Riggs, A.D.; Yu, H.; Jove, R.; Yuan, Y.C. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res* **2019**, *47*, 8335-8336, doi:10.1093/nar/gkz663.

188. Akalin, A.; Kormaksson, M.; Li, S.; Garrett-Bakelman, F.E.; Figueroa, M.E.; Melnick, A.; Mason, C.E. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **2012**, *13*, R87, doi:10.1186/gb-2012-13-10-r87.

189. Day, K.; Waite, L.L.; Thalacker-Mercer, A.; West, A.; Bamman, M.M.; Brooks, J.D.; Myers, R.M.; Absher, D. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol* **2013**, *14*, R102, doi:10.1186/gb-2013-14-9-r102.

190. Zhang, Y.; Baheti, S.; Sun, Z. Statistical method evaluation for differentially methylated CpGs in base resolution next-generation DNA sequencing data. *Brief Bioinform* **2018**, *19*, 374-386, doi:10.1093/bib/bbw133.

191. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **2014**, *15*, 550, doi:10.1186/s13059-014-0550-8.

192. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139-140, doi:10.1093/bioinformatics/btp616.

193. Rakyan, V.K.; Down, T.A.; Balding, D.J.; Beck, S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* **2011**, *12*, 529-541, doi:10.1038/nrg3000.

194. Ayyala, D.N.; Frankhouser, D.E.; Ganbat, J.O.; Marcucci, G.; Bundschuh, R.; Yan, P.; Lin, S. Statistical methods for detecting differentially methylated regions based on MethylCap-seq data. *Brief Bioinform* **2016**, *17*, 926-937, doi:10.1093/bib/bbv089.

195. Chen, D.P.; Lin, Y.C.; Fann, C.S. Methods for identifying differentially methylated regions for sequence- and array-based data. *Brief Funct Genomics* **2016**, *15*, 485-490, doi:10.1093/bfgp/elw018.

196. Tsuji, J.; Weng, Z. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Brief Bioinform* **2016**, *17*, 938-952, doi:10.1093/bib/bbv103.

197. Sun, X.; Han, Y.; Zhou, L.; Chen, E.; Lu, B.; Liu, Y.; Pan, X.; Cowley, A.W., Jr.; Liang, M.; Wu, Q.; et al. A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. *Bioinformatics* **2018**, *34*, 2715-2723, doi:10.1093/bioinformatics/bty174.

198. Yong, W.S.; Hsu, F.M.; Chen, P.Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **2016**, *9*, 26, doi:10.1186/s13072-016-0075-3.

199. Akalin, A.; Franke, V.; Vlahovicek, K.; Mason, C.E.; Schubeler, D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **2015**, *31*, 1127-1129, doi:10.1093/bioinformatics/btu775.

200. Cavalcante, R.G.; Sartor, M.A. annotatr: genomic regions in context. *Bioinformatics* **2017**, *33*, 2381-2383, doi:10.1093/bioinformatics/btx183.

201. Teschendorff, A.E.; Breeze, C.E.; Zheng, S.C.; Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *Bmc Bioinformatics* **2017**, *18*, 105, doi:10.1186/s12859-017-1511-5.

202. Kang, S.; Li, Q.; Chen, Q.; Zhou, Y.; Park, S.; Lee, G.; Grimes, B.; Krysan, K.; Yu, M.; Wang, W.; et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* **2017**, *18*, 53, doi:10.1186/s13059-017-1191-5.
203. Li, W.; Li, Q.; Kang, S.; Same, M.; Zhou, Y.; Sun, C.; Liu, C.C.; Matsuoka, L.; Sher, L.; Wong, W.H.; et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res* **2018**, *46*, e89, doi:10.1093/nar/gky423.
204. Jung, M.; Klotzek, S.; Lewandowski, M.; Fleischhacker, M.; Jung, K. Changes in concentration of DNA in serum and plasma during storage of blood samples. *Clin Chem* **2003**, *49*, 1028-1029.
205. Barault, L.; Amatu, A.; Siravegna, G.; Ponzetti, A.; Moran, S.; Cassingena, A.; Mussolin, B.; Falcomata, C.; Binder, A.M.; Cristiano, C.; et al. Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. *Gut* **2018**, *67*, 1995-2005, doi:10.1136/gutjnl-2016-313372.
206. Tian, Q.; Zou, J.; Tang, J.; Fang, Y.; Yu, Z.; Fan, S. MRCNN: a deep learning model for regression of genome-wide DNA methylation. *BMC Genomics* **2019**, *20*, 192, doi:10.1186/s12864-019-5488-5.
207. Pantel, K.; Alix-Panabieres, C. Liquid biopsy and minimal residual disease - latest advances and implications for cure. *Nat Rev Clin Oncol* **2019**, doi:10.1038/s41571-019-0187-3.
208. Gkountela, S.; Castro-Giner, F.; Szczerba, B.M.; Vetter, M.; Landin, J.; Scherrer, R.; Krol, I.; Scheidmann, M.C.; Beisel, C.; Stirnimann, C.U.; et al. Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding. *Cell* **2019**, *176*, 98-112 e114, doi:10.1016/j.cell.2018.11.046.
209. Cristiano, S.; Leal, A.; Phallen, J.; Fiksel, J.; Adleff, V.; Bruhm, D.C.; Jensen, S.O.; Medina, J.E.; Hruban, C.; White, J.R.; et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **2019**, *570*, 385-389, doi:10.1038/s41586-019-1272-6.
210. Ovcharenko, A.; Rentmeister, A. Emerging approaches for detection of methylation sites in RNA. *Open Biol* **2018**, *8*, doi:10.1098/rsob.180121.
211. Sun, K.; Jiang, P.; Chan, K.C.; Wong, J.; Cheng, Y.K.; Liang, R.H.; Chan, W.K.; Ma, E.S.; Chan, S.L.; Cheng, S.H.; et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **2015**, *112*, E5503-5512, doi:10.1073/pnas.1508736112.
212. Chan, K.C.; Jiang, P.; Chan, C.W.; Sun, K.; Wong, J.; Hui, E.P.; Chan, S.L.; Chan, W.C.; Hui, D.S.; Ng, S.S.; et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A* **2013**, *110*, 18761-18768, doi:10.1073/pnas.1313995110.
213. Su, Y.; Fang, H.B.; Jiang, F. An epigenetic classifier for early stage lung cancer. *Clin Epigenetics* **2018**, *10*, 68, doi:10.1186/s13148-018-0502-3.
214. 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*, doi:10.1126/scitranslmed.aax7533.
215. Xu, R.H.; Wei, W.; Krawczyk, M.; Wang, W.; Luo, H.; Flagg, K.; Yi, S.; Shi, W.; Quan, Q.; Li, K.; et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater* **2017**, *16*, 1155-1161, doi:10.1038/nmat4997.

216. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079, doi:10.1093/bioinformatics/btp352.
217. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **2014**, *47*, 11 12 11-34, doi:10.1002/0471250953.bi1112s47.
218. Irizarry, R.A.; Ladd-Acosta, C.; Wen, B.; Wu, Z.; Montano, C.; Onyango, P.; Cui, H.; Gabo, K.; Rongione, M.; Webster, M.; et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **2009**, *41*, 178-186, doi:10.1038/ng.298.
219. MethylCap kit (Diagenode Cat# C02020010).
220. Aberg, K.A.; Xie, L.; Chan, R.F.; Zhao, M.; Pandey, A.K.; Kumar, G.; Clark, S.L.; van den Oord, E.J. Evaluation of Methyl-Binding Domain Based Enrichment Approaches Revisited. *PLoS One* **2015**, *10*, e0132205, doi:10.1371/journal.pone.0132205.
221. Bock, C.; Tomazou, E.M.; Brinkman, A.B.; Muller, F.; Simmer, F.; Gu, H.; Jager, N.; Gnirke, A.; Stunnenberg, H.G.; Meissner, A. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* **2010**, *28*, 1106-1114, doi:10.1038/nbt.1681.
222. Huang, J.; Soupir, A.C.; Schlick, B.D.; Teng, M.; Sahin, I.H.; Permuth, J.B.; Siegel, E.M.; Manley, B.J.; Pellini, B.; Wang, L. Cancer Detection and Classification by CpG Island Hypermethylation Signatures in Plasma Cell-Free DNA. *Cancers (Basel)* **2021**, *13*, doi:10.3390/cancers13225611.
223. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **2021**, *71*, 209-249, doi:10.3322/caac.21660.
224. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **2007**, *8*, 286-298, doi:10.1038/nrg2005.
225. Warren, J.D.; Xiong, W.; Bunker, A.M.; Vaughn, C.P.; Furtado, L.V.; Roberts, W.L.; Fang, J.C.; Samowitz, W.S.; Heichman, K.A. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med* **2011**, *9*, 133, doi:10.1186/1741-7015-9-133.
226. Nian, J.; Sun, X.; Ming, S.; Yan, C.; Ma, Y.; Feng, Y.; Yang, L.; Yu, M.; Zhang, G.; Wang, X. Diagnostic Accuracy of Methylated SEPT9 for Blood-based Colorectal Cancer Detection: A Systematic Review and Meta-Analysis. *Clin Transl Gastroenterol* **2017**, *8*, e216, doi:10.1038/ctg.2016.66.
227. Sprang, M.; Paret, C.; Faber, J. CpG-Islands as Markers for Liquid Biopsies of Cancer Patients. *Cells* **2020**, *9*, doi:10.3390/cells9081820.
228. Blighe K, R.S., Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. **2021**.
229. Marini, F.; Binder, H. pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *Bmc Bioinformatics* **2019**, *20*, 331, doi:10.1186/s12859-019-2879-1.
230. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **2008**, doi:dx.doi.org/10.18637/jss.v028.i05.
231. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **2010**, *33*, 1-22.
232. Krijthe, J.H. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. **2015**.
233. Hiam-Galvez, K.J.; Allen, B.M.; Spitzer, M.H. Systemic immunity in cancer. *Nat Rev Cancer* **2021**, *21*, 345-359, doi:10.1038/s41568-021-00347-z.

234. Kalinski, P. Regulation of immune responses by prostaglandin E2. *J Immunol* **2012**, *188*, 21-28, doi:10.4049/jimmunol.1101029.

235. Arechederra, M.; Daian, F.; Yim, A.; Bazai, S.K.; Richelme, S.; Dono, R.; Saurin, A.J.; Habermann, B.H.; Maina, F. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nat Commun* **2018**, *9*, 3164, doi:10.1038/s41467-018-05550-5.

236. Bhatlekar, S.; Fields, J.Z.; Boman, B.M. Role of HOX Genes in Stem Cell Differentiation and Cancer. *Stem Cells Int* **2018**, *2018*, 3569493, doi:10.1155/2018/3569493.

237. Wei, M.; Zhang, C.; Tian, Y.; Du, X.; Wang, Q.; Zhao, H. Expression and Function of WNT6: From Development to Disease. *Front Cell Dev Biol* **2020**, *8*, 558155, doi:10.3389/fcell.2020.558155.

Appendices

Appendices 1. Copyright

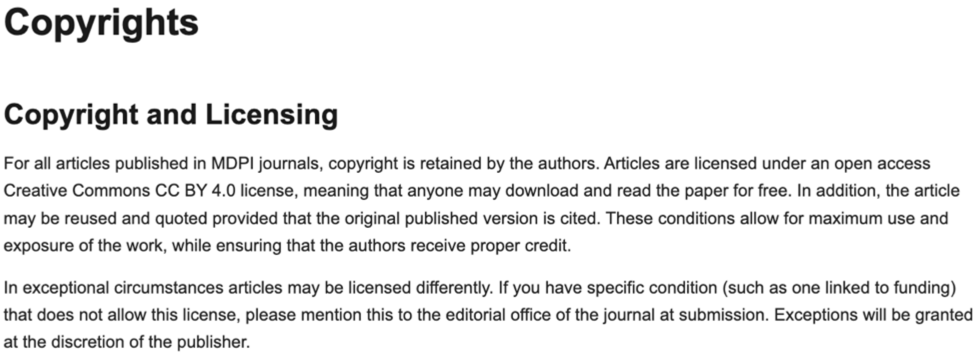
Chapter 1 and Chapter 3:

Parts of these sections were previously published by Cancers, a peer-reviewed, open access journal of oncology, published semimonthly online by MDPI.

Huang, J.; Wang, L. Cell-Free DNA Methylation Profiling Analysis - Technologies and Bioinformatics. Cancers (Basel) 2019, 11, doi:10.3390/cancers11111741.

Huang, J.; Soupir, A.C.; Schlick, B.D.; Teng, M.; Sahin, I.H.; Permuth, J.B.; Siegel, E.M.; Manley, B.J.; Pellini, B.; Wang, L. Cancer Detection and Classification by CpG Island Hypermethylation Signatures in Plasma Cell-Free DNA. Cancers (Basel) 2021, 13, doi:10.3390/cancers13225611.

According to https://www.mdpi.com/authors/rights, for all articles published in MDPI journals, copyright is retained by the authors.



Copyrights

Copyright and Licensing

For all articles published in MDPI journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.

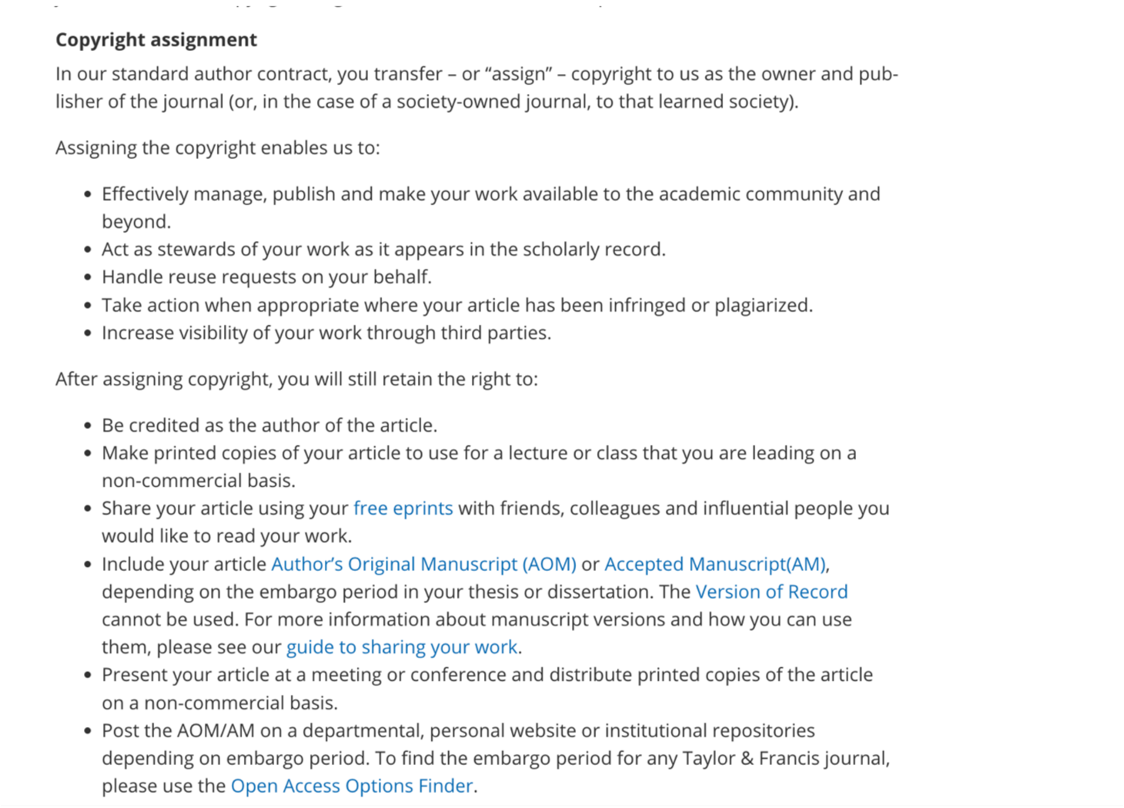
In exceptional circumstances articles may be licensed differently. If you have specific condition (such as one linked to funding) that does not allow this license, please mention this to the editorial office of the journal at submission. Exceptions will be granted at the discretion of the publisher.

Chapter 2:

Parts of this section have been accepted for publication in Epigenetics, published by Taylor & Francis.

Huang, J.; Soupir, A.C.; Wang, L. Cell-free DNA methylome profiling by MBD-seq with ultra-low input. Epigenetics 2021, 1-14, doi:10.1080/15592294.2021.1896984.

According to https://authorservices.taylorandfrancis.com/publishing-your-research/moving-through-production/copyright-for-journal-authors/, after assigning copyright, the author will still retain the right to: Include article Author's Original Manuscript (AOM) or Accepted Manuscript (AM) in a thesis or dissertation, depending on the embargo period. The embargo period of *Epigenetics* is 12 months.



Copyright assignment

In our standard author contract, you transfer – or "assign" – copyright to us as the owner and publisher of the journal (or, in the case of a society-owned journal, to that learned society).

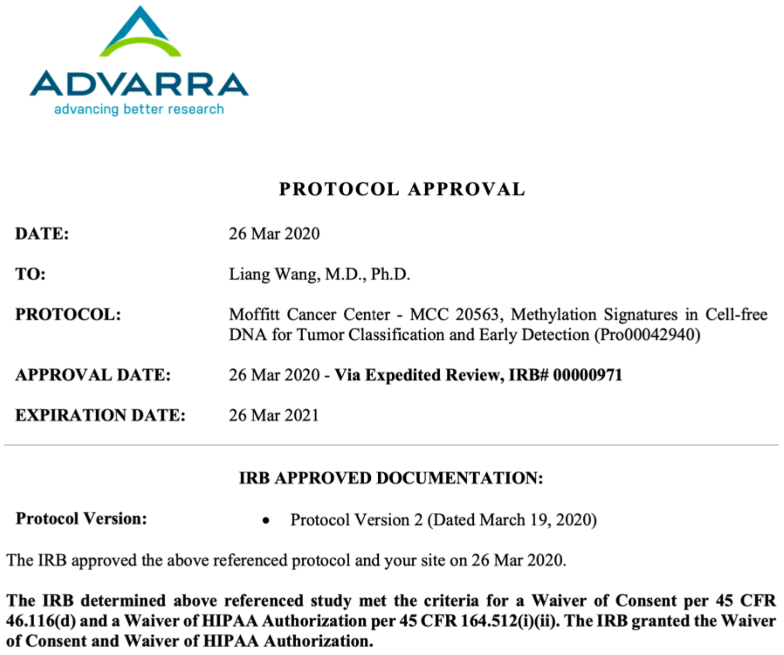
Assigning the copyright enables us to:

- Effectively manage, publish and make your work available to the academic community and beyond.
- Act as stewards of your work as it appears in the scholarly record.
- Handle reuse requests on your behalf.
- Take action when appropriate where your article has been infringed or plagiarized.
- Increase visibility of your work through third parties.

After assigning copyright, you will still retain the right to:

- Be credited as the author of the article.
- Make printed copies of your article to use for a lecture or class that you are leading on a non-commercial basis.
- Share your article using your free eprints with friends, colleagues and influential people you would like to read your work.
- Include your article Author's Original Manuscript (AOM) or Accepted Manuscript(AM), depending on the embargo period in your thesis or dissertation. The Version of Record cannot be used. For more information about manuscript versions and how you can use them, please see our guide to sharing your work.
- Present your article at a meeting or conference and distribute printed copies of the article on a non-commercial basis.
- Post the AOM/AM on a departmental, personal website or institutional repositories depending on embargo period. To find the embargo period for any Taylor & Francis journal, please use the Open Access Options Finder.

Appendices 2. IRB approvals



ADVARRA
advancing better research

PROTOCOL APPROVAL

DATE:	26 Mar 2020
TO:	Liang Wang, M.D., Ph.D.
PROTOCOL:	Moffitt Cancer Center - MCC 20563, Methylation Signatures in Cell-free DNA for Tumor Classification and Early Detection (Pro00042940)
APPROVAL DATE:	26 Mar 2020 - **Via Expedited Review, IRB# 00000971**
EXPIRATION DATE:	26 Mar 2021

IRB APPROVED DOCUMENTATION:

Protocol Version:	• Protocol Version 2 (Dated March 19, 2020)

The IRB approved the above referenced protocol and your site on 26 Mar 2020.

The IRB determined above referenced study met the criteria for a Waiver of Consent per 45 CFR 46.116(d) and a Waiver of HIPAA Authorization per 45 CFR 164.512(i)(ii). The IRB granted the Waiver of Consent and Waiver of HIPAA Authorization.

Appendices 3. Contributions

Mentor – Liang Wang: Funding acquisition and conceptualization.

Collaborative Data Services Core: Provisioning of patient-level data.

Tissue Core: Collect and release of biospecimens.

Molecular Genomics Core: Next-generation sequencing of well-prepared library.

Biostatistics and Bioinformatics Core: Power calculation.

Author – Jinyong Huang: Experiment design, cfDNA extraction, cfMBD-seq library preparation, quality control of DNA libraries, computational data analyses (quality control of sequence reads, alignment, differential methylation analyses, machine learning analyses, and data visualization), and writing.