UNIVERSITY OF CALIFORNIA, SAN DIEGO

Novel molecular and bioinformatics approaches to investigate DNA methylation in human epigenome

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor in Philosophy

in

Bioinformatics and Systems Biology

by

Dinh Hue Diep

Committee in charge:

Professor Kun Zhang, Chair
Professor Vineet Bafna, Co-Chair
Professor John Chang
Professor Trey Ideker
Professor Bing Ren
Professor Sheng Zhong

2017

The Dissertation of Dinh Hue Diep is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-chair

_____
Chair

University of California, San Diego

2017

DEDICATION

To Mẹ, Ba, Phương, Linh, Phong, and Richard.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

5caC	5-carboxylmethylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AMF	Average methylation frequency
BSPP	Bisulfite padlock probes
BS-seq	Bisulfite sequencing
caHMH	Cancer associated high methylation haplotype
cfDNA	cell free DNA
ctDNA	circulating tumor DNA
DHS	DNase hypersensitive site
DMR	Differentially methylated region
DMS	Differentially methylated site
ENCODE	Encyclopedia of DNA Elements
FDR	False discovery rate
FDR	False discovery rate
FET	Fisher’s exact test
hMHB	Hydroxymethylation haplotype block
IMF	Individual methylation frequency
JSD	Jensen-Shannon divergence
LAD	Lamin associated domains
LD	Linkage disequilibrium
LOCK	Large organized chromatin Lys9 modifications
MHB	Methylation haplotype block
MHL	Methylation haplotype load
NGS	Next generation sequencing
oxBS-seq	Oxidative bisulfite sequencing
PCR	Polymerase chain reaction

RAM	Random access memory
RMSE	Root mean square error
RRBS	Reduced representation bisulfite sequencing
sc-RRBS	Single cell RRBS
SNP	Single nucleotide polymorphism
SRA	Short reads archive
TAB-seq	Tet-assisted bisulfite sequencing
TAD	Topological associated domains
TF	Transcription factor
TFBS	Transcription factor binding site
UMI	Unique molecular identifier
VMR	Variably methylated region
WB	Whole blood
WBC	White blood cell
WGBS	Whole genome bisulfite sequencing

ACKNOWLEDGEMENTS

I thank my advisor Kun Zhang for his advice and guidance throughout the years. I will always appreciate his contribution to my progress as a researcher.

I am also thankful to have the attention and advice of my committee, Vineet Bafna, John Chang, Trey Ideker, Bing Ren, and Sheng Zhong.

Throughout my PhD, I was fortunate to work with my many colleagues. I am indebted to my co-authors for their contributions to this dissertation. My knowledge of experimental molecular biology benefited from working alongside Jie Deng, Sam Chiang, Zhe Li, Ho-Lim Fung, and Nongluk Plongthongkum. I also learned much of high throughput sequencing analysis from Athurva Gore and DNA methylation analysis from Shicheng Guo. Many wonderful people from the Zhang lab have also been very helpful to me. I would especially like to thank Blue Lake and Rui Liu for being always available to share their knowledge with me.

I am thankful to Sergio Ruiz, Shen Li, Sarah Tiskhoff, Tiffany Tanaka, and Dana Tsui for our close collaborations and providing me an opportunity to explore new research interests.

I am fortunate to be in the Bioinformatics Program and be surrounded by many interesting and friendly folks throughout the years. I especially thank my friends Daria Merkurjev, Shamim Mollah, Joanne Liu, and Anugraha Raman for many great dinners.

I also thank the CIRM stem cell training grant for providing me with two years of generous support and training, during which I was able to attend many exciting meetings and conferences.

Lastly, I would like to thank my family and Richard Que for being there for me every step of this journey.

Chapter 2, contains material as it appears in: Dinh Diep*, Nongluk Plongthongkum*, Athurva Gore*, Ho-Lim Fung, Robert Shoemaker, Kun Zhang. “Library-free Methylation Sequencing with Bisulfite Padlock Probes.” Nature Methods. 2012 February 5; 9(3): 270-272. doi: 10.1038/nmeth.1871. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3, contains material from a submitted manuscript: Dinh Diep and Kun Zhang. “*cgDMR-miner*: generalized method for the identification of differentially methylated regions on shallow WGBS datasets”. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4, contains material as it appears in: Guo, Shicheng*, Dinh Diep*, Nongluk Plongthongkum, Ho Lim Fung, Kang Zhang, Kun Zhang. “Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA”. Nature Genetics. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2009 Bachelor of Science in Bioengineering, Biotechnology, University of California, San Diego

2017 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California, San Diego

Publications

1. Guo, Shicheng*, **Dinh Diep***, Nongluk Plongthongkum, Ho Lim Fung, Kang Zhang, Kun Zhang. “Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA”. Nature Genetics (March 6, 2017): [Epub ahead of print]. doi: 10.1038/ng.3805.
2. BLUEPRINT consortium. “Quantitative Comparison of DNA Methylation Assays for Biomarker Development and Clinical Applications.” Nature Biotechnology 34, no. 7 (July 2016): 726–37. doi:10.1038/nbt.3605.
3. Plongthongkum, Nongluk*, **Dinh H. Diep***, and Kun Zhang. “Advances in the Profiling of DNA Modifications: Cytosine Methylation and Beyond.” Nature Reviews Genetics 15, no. 10 (October 2014): 647–61. doi:10.1038/nrg3772.
4. Shen, Li, Hao Wu, **Dinh Diep**, Shinpei Yamaguchi, Ana C. D’Alessio, Ho-Lim Fung, Kun Zhang, and Yi Zhang. “Genome-Wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics.” Cell 153, no. 3 (April 25, 2013). doi:10.1016/j.cell.2013.04.002.
5. Yamaguchi, Shinpei, Kwonho Hong, Rui Liu, Li Shen, Azusa Inoue, **Dinh Diep**, Kun Zhang, and Yi Zhang. “Tet1 Controls Meiosis by Regulating Meiotic Gene

Expression." Nature 492, no. 7429 (December 20, 2012): 443–47. doi:10.1038/nature11709.

6. Ruiz, Sergio*, **Dinh Diep***, Athurva Gore, Athanasia D. Panopoulos, Nuria Montserrat, Nongluk Plongthongkum, Sachin Kumar, et al. "Identification of a Specific Reprogramming-Associated Epigenetic Signature in Human Induced Pluripotent Stem Cells." Proceedings of the National Academy of Sciences of the United States of America 109, no. 40 (October 2, 2012): 16196–201. doi:10.1073/pnas.1202352109.
7. **Dinh Diep***, Nongluk Plongthongkum*, Athurva Gore, Ho-Lim Fung, Robert Shoemaker, and Kun Zhang. "Library-Free Methylation Sequencing with Bisulfite Padlock Probes." Nature Methods 9, no. 3 (March 2012): 270–72. doi:10.1038/nmeth.1871.
8. Thiele, Ines, Ronan M. T. Fleming, Richard Que, Aarash Bordbar, **Dinh Diep**, and Bernhard O. Palsson. "Multiscale Modeling of Metabolism and Macromolecular Synthesis in E. Coli and Its Application to the Evolution of Codon Usage." PloS One 7, no. 9 (2012). doi:10.1371/journal.pone.0045635.
9. Panopoulos, Athanasia D., Oscar Yanes, Sergio Ruiz, Yasuyuki S. Kida, **Dinh Diep**, Ralf Tautenhahn, Aida Herrerias, et al. "The Metabolome of Induced Pluripotent Stem Cells Reveals Metabolic Changes Occurring in Somatic Cell Reprogramming." Cell Research 22, no. 1 (January 2012): 168–77. doi:10.1038/cr.2011.177.
10. Kasper Hansen, Winston Timp, Hector Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G. McDonald, Bo Wen, et al. "Increased Methylation Variation in Epigenetic Domains across Cancer Types." Nature Genetics 43, no. 8 (August 2011): 768–75. doi:10.1038/ng.865.

11. **Dinh Diep**, and Kun Zhang. “Genome-Wide Mapping of the Sixth Base.” Genome Biology 12, no. 6 (June 20, 2011). doi:10.1186/gb-2011-12-6-116.

12. Liu, Guang-Hui, Basam Z. Barkho, Sergio Ruiz, **Dinh Diep**, Jing Qu, Sheng-Lian Yang, Athanasia D. Panopoulos, et al. “Recapitulation of Premature Ageing with iPSCs from Hutchinson-Gilford Progeria Syndrome.” Nature 472, no. 7342 (April 14, 2011): 221–25. doi:10.1038/nature09879.

ABSTRACT OF THE DISSERTATION

Novel molecular and bioinformatics approaches to investigate DNA methylation in human epigenome

by

Dinh Hue Diep

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Kun Zhang, Chair

Professor Vineet Bafna, Co-Chair

In recent years, advances in sequencing have enabled the mapping of DNA methylation variation across many different populations of human cell types and the identification of candidate DNA methylation biomarkers for clinical applications. The clinical development of DNA methylation biomarkers has been limited, however, due to the high cost and the lack of flexibility in using the current experimental and bioinformatics tools.

We made improvements to the design of bisulfite padlock probes (BSPP) to greatly increase efficiency and throughput for targeted DNA methylation quantification. The cost effectiveness and scalability of this approach was demonstrated on hundreds of samples using a set of 330,000 probes. We also developed a bioinformatics pipeline that performs SNP calling on bisulfite data and DNA methylation quantitation with reduced errors from various different assay types.

Despite many available bioinformatics tools for differential DNA methylation analysis, there is a need for a more general computational tool to characterize DNA methylation variability on reference data. Therefore, we developed a new differential methylation identification method and variability score to quantify DNA methylation variation across multiple groups of samples. For simulated 5X average depth of coverage datasets, *cgDMR-miner*, identified 42% of simulated DMRs with 73% precision while the next best approach identified 23% of simulated DMRs with 96% precision. Thus *cgDMR-miner* can identify potential targets from a shallow, low accuracy initial screen that can later be validated with a deeper screen using a targeted assay.

Lastly, the coordinated methylation of nearby CpG sites was investigated in order to identify more robust biomarkers for cancer. Starting with a set of identified 147,888 regions of tightly coupled CpG methylation or methylation haplotype blocks (MHBs), the linked status of CpGs within these regions were found useful for biomarker identification in human tissue samples and human cell free DNA.

Chapter 1: Introduction

Human epigenomes contain a plethora of chemical compounds that mark DNA and proteins which attach to DNA to orchestrate regulation of genes expression and cellular activities. Most importantly, they enable diversity of cells and tissue types in multicellular organisms. Epigenetic phenomena are plastic and often susceptible to the environmental and behavioral influences[1–3]. In early mammalian development, epigenetic marks are erased and then reestablished after fertilization and also in the development of primordial germ cells. Transgenerational epigenetic inheritance has only been shown for a few cases with the strongest evidence found in plants and some animal species[4]. While initiation of epigenetic phenomena are established by epigenetic 'writers', inheritance between cell divisions may be carried out by alternative mechanisms which are thought to be guided by cell memory[5]. Thus, identification of differential patterns in the epigenome could generate a biomolecular roadmap to disease pathogenesis.

Two well-studied epigenetic phenomena are histone modification and DNA methylation. Histones are proteins that comprise the basic units of nucleosomes. These nucleosomes are then utilized in the packaging of DNA within the nucleus. The post-translational modification of histones such as methylation and acetylation, can influence the 3-D structure of the nucleus and modulate gene expression patterns. DNA methylation is a stable and reversible epigenetic mark that occurs on the fifth carbon of cytosines and chemically named "5-methylcytosine" (5mC). It can be reversed passively over mitotic divisions or reversed actively via enzymatic processing and oxidation. Many important biological processes are characterized by DNA methylation, such as X-chromosome inactivation in females, transcriptional repression of transposons, genomic imprinting, and alleles specific silencing of genes in different tissues[6,7]. Oxidized derivatives of 5mC, 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-

carboxylcytosine (5caC), may also carry out important functions in embryonic development and notably have been linked with cancer development[8–11].

In this work, we focus on developing experimental techniques for profiling DNA methylation and novel bioinformatics approaches in the quantification of DNA methylation differences with applications towards biomarkers development. Also, we focus on methods that are compatible with next-generation sequencing (NGS) and capable of absolute measurements. In the next section, we describe DNA methylation in human epigenomes. The following sections give an overview of the conventional single base resolution quantification methods used in the study of 5mC and 5hmC. The final section will give an overview to differential DNA methylation analysis from next-generation sequencing datasets.

DNA methylation in human epigenomes

DNA methylation in human cells is generally at CpG dinucleotides. Human genomes are highly methylated throughout with pockets of high density CpG regions known as CpG islands that are mostly unmethylated. While nearly absence in almost all somatic cell types, non-CpG methylation patterns have been associated with brain development, pluripotency, and diseases such as Rett syndrome and diabetes[12]. In this work, we focus on DNA methylation at CpG sites only, owing to the limited understanding of non-CpG methylation in human currently. Methylated CpGs are able to mutate to TpG or CpA through deamination of the methylated cytosine, so the existence of high density CpG regions or CpG islands are indicative of some selection pressure in human genomes to maintain these regions and may explain the lack of methylation in these regions. Dynamic DNA methylation is often observed at promoters with low CpG density sequences which means that other promoters with high CpG density sequences tend be regulated via alternative epigenetic mechanisms[13]. It also shown that in normal

development, cells may switch between DNA methylation and alternative mechanisms for regulating genes expression[13].

DNA methylation writers include the DNA methyltransferases: DNMT1, DNMT3A, and DNMT3B. DNMT1 ensures that 5mC is stably maintained through mitotic divisions while DNMT3A/B generates *de novo* 5mC marks. The enzymes that catalyze the reactions leading to de-methylation belong to the ten-eleven-translocation (TET) protein family and include TET1, TET2, and TET3[14]. These proteins catalyze the oxidation of 5mC to 5-hydroxymethylcytosines (5hmC). Further sequential oxidations lead to 5-formylcytosine (5fC) and then 5-carboxylcytosine (5caC). Human thymine DNA glycosylase (TDG) can excise 5fC and 5caC to generate an abasic site that goes on to activate the base excision repair pathways, resulting in an unmodified cytosine[15,16]. TET protein mediated 5mC turnover is a significant phenomenon that occurs both during embryogenesis and the formation of primordial germ cells[17]. The intermediates of de-methylation, 5hmC, 5fC and 5caC, are sometimes stably maintained and perform biological functions such as epigenetic priming and modulating transcriptional timing[18–25].

DNA methylation could inform clinical prognosis of patients and their response to drugs, including methylation inhibiting drugs, over time[26–28]. Additionally, minimally invasive testing could be performed on circulating cell free DNA from cancer patients[29–31]. Further developments of DNA methylation quantification technologies and bioinformatics, however, are needed to enable clinical tools utilizing DNA methylation as biomarkers.

Conventional DNA methylation quantification approaches

The most widely used method for mapping DNA methylation at a single base pair resolution involves sequencing DNA after a chemical conversion of unmethylated cytosines to uracils known as bisulfite sequencing (BS-seq). The sequencing readouts

will show thymidines (T) at cytosine positions which were unmethylated due to uracils-adenine pairing during polymerase chain reaction (PCR). Recent discoveries of quantifiable amounts of 5hmC in various mammalian cells (up to 1% of cytosines), especially in embryonic and brain tissues, have also led to a recognition that 5hmC is resistant to deamination like 5mC in bisulfite treatment[32]. Nonetheless, as 5hmC levels are typically low in most tissues types, BS-seq is still widely used for DNA methylation profiling. Another popular methodology for genome-wide mapping 5mC is reduced representation bisulfite sequencing (RRBS)[33]. While WGBS is considered the most comprehensive method for DNA methylation mapping, it is currently unsustainable for studies with large sample sizes because of the high sequencing cost. In bulk library preparations using these approaches, an average measurement across a population of cells is taken and therefore noisy sampling and bias from various technical aspects of library preparation all can introduce errors in methylation quantification. Thus, whole genome methylation quantification typically requires higher sampling or higher depth of coverage than whole genome sequencing (typically 30X versus 3X). RRBS, though more cost-effective, leaves many important locus such as enhancers within low-CpG density regions uncharacterized[34]. While both WGBS and RRBS are good candidates for biomarkers discovery, due to high cost of WBGS and inflexibility of RRBS, they have limited use for clinical applications and biomarker development.

Conventional DNA hydroxymethylation quantification approaches

Quantification of 5-hydroxymethylcytosines has been largely performed using non-targeted enrichment approaches or in combination with arrays because the levels of 5hmC are estimated to be less than 1% of total cytosine in human and mouse cells[16]. Non-targeted enrichment methods utilizes either specific antibodies, 5hmC sensitive restriction enzymes, or chemical labelling to enrich for 5hmC carrying DNA fragments[35].

Whole genome methods to quantify 5hmC have been developed utilizing enzymatic or chemical treatments in addition to bisulfite conversion. The first of these methods is Tet-assisted bisulfite sequencing (TAB-seq)[36] which provides absolute quantification of 5hmC. The second method capable of 5hmC quantification at single base resolution is oxidative bisulfite sequencing (oxBS-seq)[37]. Sequencing readouts for oxBS-seq will generate cytosines at 5mC positions and thymidines at non-5mC positions. Subtraction of oxBS-seq (5mC only) profiles from BS-seq (5mC+5hmC) profiles results in 5hmC only profiles[38]. The DNA shearing step in both TAB-seq and oxBS-seq can be replaced with *Msp*I digestion, which results in reduced representation libraries[35,39]. The final library from both TAB-seq and oxBS-seq can also be applied to Infinium arrays, which can provide base resolution quantification of targeted CpG sites[8,10,40,41].

Differential DNA methylation analysis

Differential DNA methylation analysis is performed on two or more bulk samples and can be challenging due to sampling noise and various technical artifacts from the experiments. A bulk sample is a sample with genomic content from a population of cells. In this work, we are mainly focused on absolute DNA methylation quantification, which is the estimated number of methylated cytosines at a single CpG position for a population of cells. Assuming only sample noise as the source of error, differential methylation analysis can be done with the Fisher's exact test (FET) for two bulk samples[42,43]. When biological variations or more than one sample is available from each group, the beta-binomial model is more commonly used [DSS[44], MOABS[45], RADMeth[46], MethylSig[47]]. In the beta-binomial framework, the observed methylation counts are binomial distributed while the methylation proportions are varied according to a beta distribution. Alternative methods utilizes the binomial distribution within a logistic regression framework [MethylKit[48]], root mean square test [MethylPy[49]], local likelihood estimation and T-test

[Bsmooth[50]], Wilcoxon or Krushal-Wallis paired non-parametric tests [MethylPipe[51]], wavelet-based functional mixed models [WFMM[52]], and full Bayesian partition model [MethyBayes[53]]. These methods have various tradeoffs between the ability to test different experimental settings, ability to determine regions and ability to account for covariates.

Chapter 2: Library-free bisulfite padlock probes

Introduction

In 2009, Deng et al developed bisulfite padlock probes (BSPP) to assay DNA methylation at targeted regions reliably in dozens of human samples[54]. The unique advantages of BSPP over conventional methods are that every locus can be potentially captured by padlock probes, and that the assay can be performed on small or large genome-wide target sets.

The design of padlock probes consists of a common linker sequence that connects two capture arms that can hybridize to two neighboring genomic regions. The probes anneal to genomic regions that have been chemically modified by bisulfite. The first capture arm from the 3' end anneals in the forward direction, and the linker sequence, which is non-complementary, provides space to "invert" and allows the second capture arm to anneal in the reverse direction to a region upstream of the first capture arm. The gap in between the two arms are "captured" by extension of with thermostable polymerase from 5' of the second capture arm to the 3' end of the first, without displacing or degrading it. Thus, the polymerase used in this reaction must not have strand displacement or 5' exonuclease activity. After extension, a thermostable ligase is used to anneal the extension end to the 3' end of the probe, creating a circular DNA product. The circular DNA product is then amplified using PCR primers that anneal to the common linker sequence.

The primary challenge in bisulfite padlock probe capture is off-target annealing. However, this is dealt with experimentally and computationally. Bisulfite converted sequences have very few cytosines so the annealing arms cannot have G's. Low CG content leads to low melting temperatures of DNA sequences. Annealing arms are then made 25-30 bp long to obtain melting temperatures above 60 degrees centigrade.

Additionally, polymerase extension requires exact matching near the 5' end of the probe and near the 3' end of the probe for ligation. Computationally, the two arms can be designed to not perfectly anneal elsewhere, although this still leaves the possibility for partial annealing of either arms to off-target regions. Off-target annealing of individual capture arms may not be captured because of polymerase inability to extend or ligase inability to create circular products, but may still cause probes to be inefficient. The second challenge in bisulfite probe design is avoiding variations such as polymorphisms and CG sites that may be methylated. If CG sites within capture arms cannot be avoided, then multiple versions of the probe should be made to allow annealing to both methylated and unmethylated sites.

BSPP capture is a flexible, high-throughput and low cost method that can be applied towards both discovery and validation of biomarkers. Additional technological improvements to reduce the time and cost of sample preparation, data quality, and comprehensive data interpretation in the form of a bioinformatics pipeline to analyze DNA methylation data at many scales would facilitate the use of DNA methylation biomarkers in the clinics. Some of these improvements have contributed to a number of recent studies on mouse and human cells [42,43,55–58].

In this chapter, we report the technical details of an improved second-generation bisulfite padlock probe (BSPP) capture method for targeted DNA methylation analysis. Specifically, we developed a new library-free protocol that dramatically reduces the time and cost of sample preparation and is compatible with automation, and a bioinformatics pipeline calls BisReadMapper that accurately and efficiently obtains both methylation levels and SNP genotypes from targeted or whole genome bisulfite sequencing data.

Results

Development of BS-seq bioinformatics pipeline

A bottleneck in bisulfite sequencing is a lack of computational tools to efficiently analyze sequencing data generated from hundreds of samples. To overcome this issue, we developed a bisulfite sequence analysis pipeline for bisulfite read mapping and DNA methylation quantification called BisReadMapper (**Figure 2-1**). BisReadMapper determines the origin strand of the read based on base composition and maps reads as if they were fully bisulfite-converted to a fully bisulfite-converted genome sequence, allowing mapping of both bi-directional and uni-directional bisulfite libraries in an unbiased manner. Another new feature is the capability to call single nucleotide polymorphisms from bisulfite sequencing data; this feature not only allows for the added analysis of allele-specific methylation[59], but also allows samples to be easily tracked in large-scale experiments.

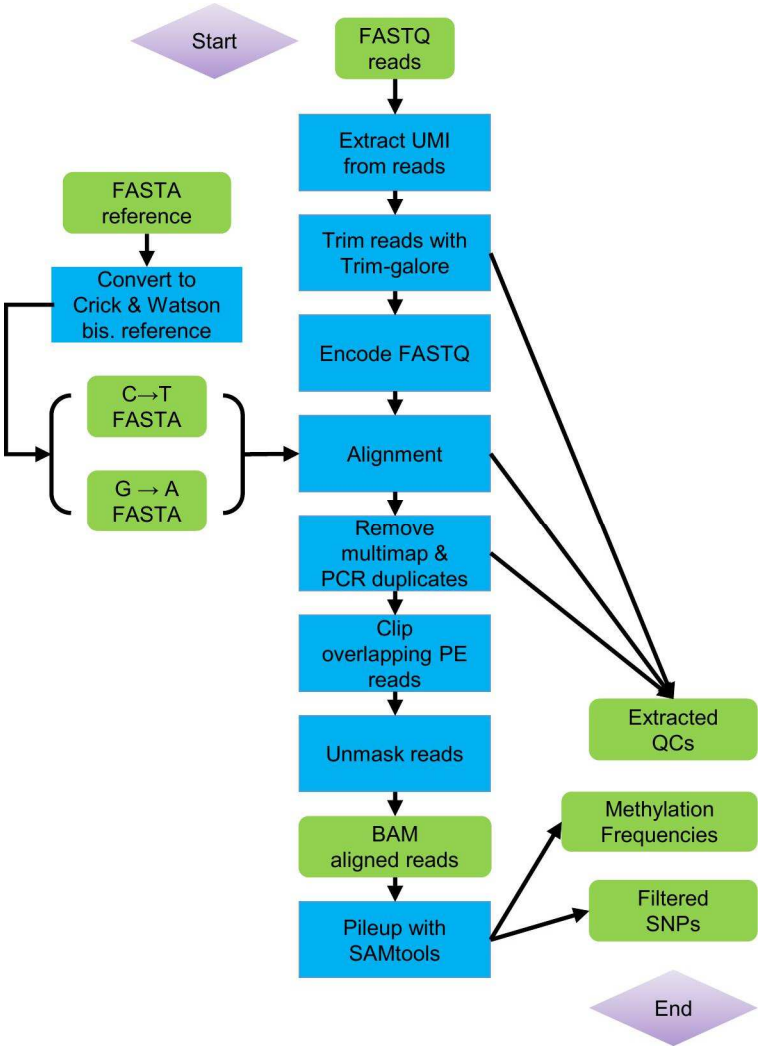


Figure 2-1. Flowchart for bisulfite reads analysis pipeline, BisReadMapper v1.4

Finally, BisReadMapper applies adaptor trimming and reads trimming rules that are specific to library types, either whole genome, reduced representation, or bisulfite padlock capture libraries. Post-alignment, BisReadMapper also performs clipping of overlapping pair-end reads. Clipping overlapping pair-ends and trimming are important for accurate quantification of methylation calls from current NGS technologies. BisReadMapper was written in Perl and runs on any Unix-based operating system. It can process 1 Gb of bisulfite-converted single-end or paired-end raw sequencing reads in less than one hour on a computational node with 4 CPU cores. The run time grows linearly as the amount of raw sequencing data increases to the typical data size of whole genome bisulfite sequencing experiments (~70 – 150 Gb).

Currently, the aligners that can be used with BisReadMapper are bowtie2[60] and BWA mem[61]. When the relative performances for each aligner on real data are compared, we found that the actual competitive advantage of each mapper depends on read length, sequencing quality, and data sizes. By utilizing the prior knowledge that read 1 are always from the reverse complementary orientation and read 2 are always from the forward orientation, we can count the number of times each mapper assigns the wrong orientation to each read. This provides a lower bound rough estimate of mapping error, which appears to be negligible, but shows the relative accuracies between aligners (**Table 2-1**). BisReadMapper can also call variants with high confident at known variant positions. When we used BisReadMapper to call variants on BSPP experiment data, we obtained 95-98% agreement with calls made by the Illumina 1M Duo bead chip (Data not shown).

Table 2-1. Performance comparison of different aligners

Aligner	Mapping rate (%)	Time (min)	Max mem. usage (GB)	Incorrect mappings	CPU threads	Dataset[1]
Bowtie2	93.9	52	6.1	161	8	A
	73.4	30	6.1	261	8	B
BWA mem	92.5	50	7.8	59	8	A
	71.7	43	7.8	5	8	B

[1]Dataset A is 3.3M quality trimmed PE 124bp reads. Dataset B is 1.6M quality trimmed PE 150bp reads.

Development of library-free BSPP

Key requirements for methylation analysis of large sample sizes include low cost, simple workflow and automation compatibility. As the cost of DNA sequencing has rapidly decreased, sample processing has become a bottleneck in terms of cost and throughput. A complicated workflow increases variability between samples and reduces power in large-scale studies. To address these issues, we extended a 'library-free' protocol[62] to multiplexed BSPP capture (**Figure 2-2**). This method eliminates five steps from Illumina's library-construction protocol such that multiplexed libraries can be generated from DNA in only four steps (**Table 2-2**). Using multiplexed primers with 6–base pair (bp) barcodes, we have routinely generated libraries for 96 samples in 96-well plates and sequenced all at once in a single Illumina HiSeq flowcell. Additionally we designed barcodes to process 384 samples per batch. As sample-specific barcodes were added, barcoded libraries can be pooled for size selection, which is the most time consuming, contamination-prone and error-prone step if performed individually. The protocol is compatible with the use of multichannel pipettes or liquid-handling devices. It dramatically reduced experimental cost and time, and improved reproducibility and read mapping rates. For large sample sizes, the library preparation cost (including probes) with our protocol was comparable to that of the reduced-representation bisulfite

sequencing and whole-genome bisulfite sequencing protocols, and the sequencing cost was much lower than that of whole-genome bisulfite sequencing owing to targeting of CpG sites of interest. Reduced-representation bisulfite sequencing is more cost-effective than BSPPs, but the former lacks BSPPs' flexibility in selecting specific sites or regions.

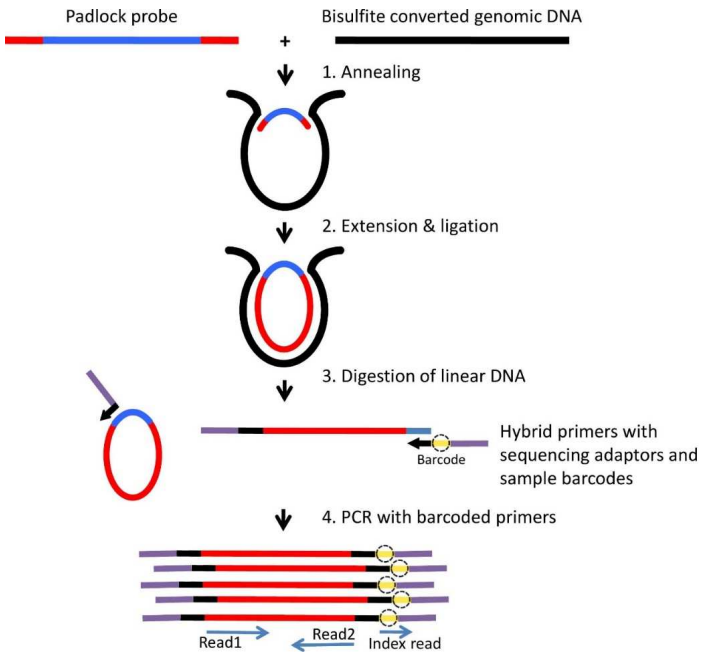


Figure 2-2. Schematic of capture experiment

Schematic of library-free BSPP protocol. Each padlock probe had a common linker sequence flanked by two target-specific capturing arms (red) that annealed to bisulfite converted genomic DNA (black), and the 3' end was extended and ligated with the 5' end to form circularized DNA. After removal of linear DNA by exonucleases, all circularized captured targets were PCR-amplified with barcoded primers. These amplicons can be directly sequenced with an Illumina sequencing platform (GA II(x) or HiSeq). Amplicon size is 363 bp, which includes captured target (180 bp), capturing arms (55 bp), and amplification primers and adapters (128 bp). The inserts can be read through with paired-end 120 bp sequencing reads.

Table 2-2. Comparison of bisulfite sequencing methods

	Published BSPP	N2-adapter BSPP	Library-free BSPP	RRBS	WGBS
Enzymatic reactions	10	6	3	4	3
Purification	6	4	1	3	3
Size-selection	2	1	1[1]	1	1
Cost per sample	$71.15[1]	$58.23[2]	$37.86[2]	$28.15	$31.10
Mapping rate	44%	80%	87%	27%[3]	N.D.
Genome coverage obtained at 10x depth	<0.1%	0.6%-1%	0.6%-1%	~1%[3]	76-96%[4]
Sequencing (Gbps)	0.5	3.2	4.0	1.4	70.0
Sequencing cost per sample[5]	$24.38	$156.00	$195.00	$68.25	$3412.50

[1]BSPP protocol size selection is typically performed on 48-96 pooled libraries. [2] Includes the cost of ordering 400K synthesized probes from LC Sciences and reagents for preparing probes, bisulfite conversion, capture, and sequencing library preparation. Estimates assume that 10K samples will be processed. [3] Estimated from: Gu et. al., Nat Methods 2010; 7(2):133-136.[4] Adapted from: Beck et. al., Nat Biotechnol 2010;28:1026-1028. [5] Assumes sequencing using an Illumina HiSeq to generate 300 Gbps of sequencing data, with cost of $4920 for a flowcell, $6815 for sequencing reagents, and $2890 for service fee. ($48.75 per Gbps)

Design and development of DMR220K probe set

To test our assay, we generated a genome-scale probe set based on our previous results and new information about differential methylation[54,63–65]. We targeted our new design for evaluation of methylation at genomic locations known to contain differentially methylated regions or differentially methylated sites[63–66], transcriptional repressor CTCF binding sites and DNase I–hypersensitive regions. We also targeted all microRNA genes and all promoters for human US National Center for Biotechnology Information reference sequence (RefSeq) genes. Using *ppDesigner* [http://genome-tech.ucsd.edu/public/Gen2_BSPP], we designed ~330,000 padlock probes that covered 140,749 non-overlapping regions with a total size of 34 megabases. We performed capturing experiments and end-sequencing, and found that these probes were slightly more specific (~96% on-target) and uniform than previous probes (**Figure 2-3**). To improve uniformity, we normalized the experimental capturing performance of these probes using subsetting and suppressor oligonucleotides as described previously[54]. We

could characterize roughly 500,000 CpG sites with ~4 gigabases of sequencing reads, and additional sites became callable with deeper sequencing.

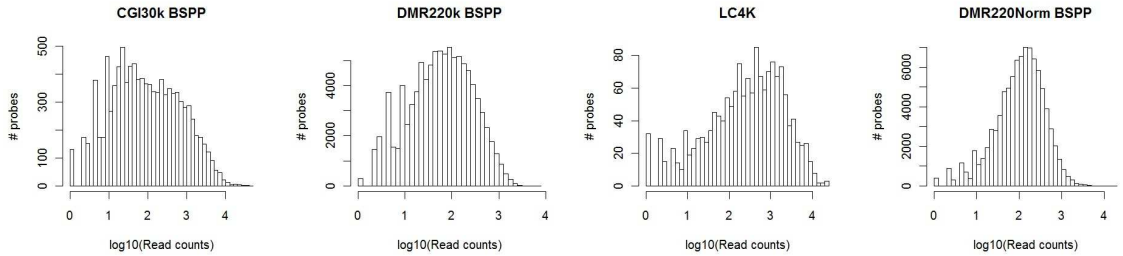


Figure 2-3. Comparison of probe capture efficiencies between the DMR220K, LC4K probe sets and the previously published CGI30K set.
The first three plots (CGI30k BSPP, DMR220k BSPP, LC4K) were generated from data without subsetting or suppressor oligos to allow for a direct comparison of probe design.

We used these probes to analyze H1 embryonic stem cells (H1 ESCs), PGP1 fibroblasts and two technical replicates of PGP1 fibroblast–derived induced pluripotent stem cells (PGP1-iPSCs). For each sample, we sequenced on average ~3.66 gigabases and measured methylation for an average of 480,904 CpG sites. To assess whether these data could be used to identify potential epigenetic regulation of transcription, we used the genomic regions enrichment of annotations tool[67] to predict the cis-regulatory potential of regions around captured CpG sites. In total, the padlock probes captured CpG sites in regions predicted to regulate 98% of RefSeq genes (**Figure 2-4**).

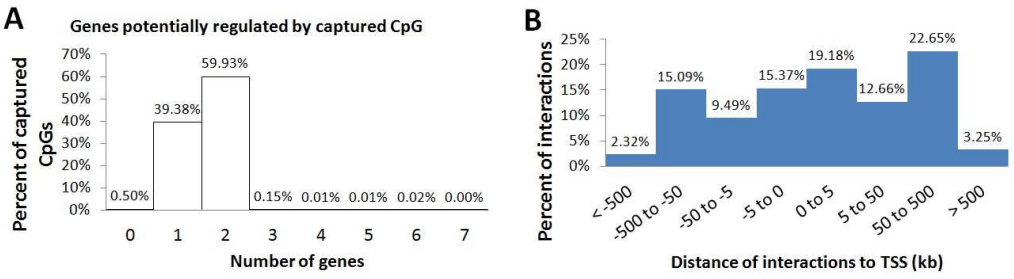


Figure 2-4. Distribution of CpGs captured wrt nearby genes
Captured CpG sites were tested for potential regulatory interactions with genes by GREAT (http://great.stanford.edu). **(A)** Most CpG sites were interacting with 1-2 genes. (B) Distance of CpG sites to the transcriptional start sites (TSS) of the predicted regulating genes.

The data generated with BSPPs accurately represented the methylation status of the target regions. Methylation levels for the two technical replicates of PGP1-iPSCs were consistent both within a single batch and between separate batches (Pearson's correlation coefficient R = 0.97–0.98). Additionally, when we compared methylation levels between technical replicates, no CpG site was different by a Fisher Exact Test with Benjamini-Hochberg multiple testing correction (false discovery rate = 0.01, n = 439,090). In comparison, large fractions of sites were differentially methylated owing to either the process of nuclear reprogramming (27.9% DMSs between PGP1-iPSCs and PGP1 fibroblasts) or the difference in cell type (31.3% DMSs between PGP1 fibroblasts and H1 ESCs) with the same criteria (false discovery rate = 0.01, n = 444,111 and 359,290, respectively). Our BSPP results with H1 ESCs were consistent with the published whole-genome sequencing of bisulfite-converted DNA[65] and published Infinium HM450K data[68] (**Figure 2-5**).

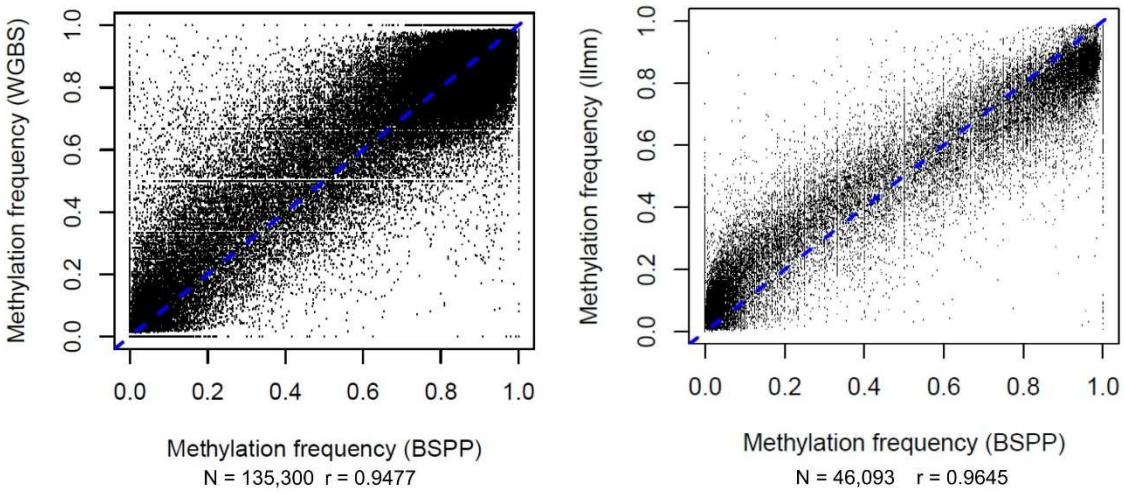


Figure 2-5. Comparison of BSPP with WGBS and Infinium HM450K
Comparison between BSPP and whole genome bisulfite sequencing (WGBS) and with Infinium HM450K (Ilmn). We compared H1 ESC datasets which are from different cultures from different labs using sites with at least 10x read depth in each.

Our assay has very low technical variability. We performed the assay on over 150 samples in 96-well plates; the yield for each was similar (**Figure 2-6**). Approximately 10% of CpG sites were targeted separately on each strand, allowing low-quality datasets with poor correlation between these built-in technical replicates to be identified. As our BSPP assay measures absolute methylation, no normalization is necessary as long as the internal replicates are consistent. Therefore, many datasets, even those generated in different laboratories, can be directly compared without batch effects, which is important for case-control studies on large samples or for meta-analyses. Additionally, the SNP-calling feature of BisReadMapper allowed us to characterize roughly 20,000 SNPs for each sample with an accuracy of 96% or better. This allowed us to unambiguously track samples, which is crucial for projects involving large sample sizes.

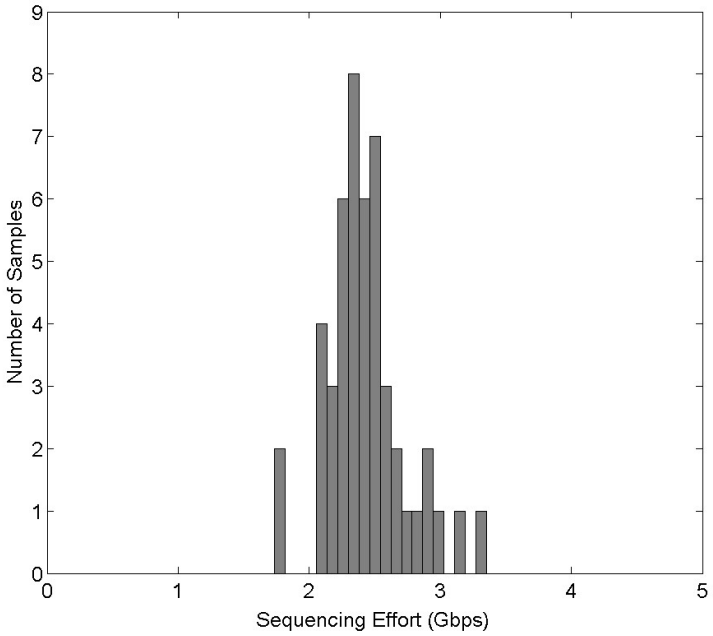


Figure 2-6. Distribution of sequencing effort per sample
Variation in amount of sequencing data obtained per sample in a multiplex BSPP capture experiment. 48 whole blood samples were captured and sequenced in one batch using the library-free BSPP method. There is little variation between samples in the amount of generated sequencing data.

Our library-free BSPP method is flexible for different study designs (**Table 2-3**). Whereas our genome-scale probe set allows global profiling on thousands of samples, a focused assay is often necessary to follow up on tens to hundreds of candidate regions identified in genome-scale scanning. Such an assay needs to be customizable to different genomic targets, scalable to a very large sample size (1,000–100,000 samples), and inexpensive.

Table 2-3. Representative cost per sample

Expected number of samples to be processed	Probe set sizes		
	4,000	**40,000**	**400,000**
Library-free protocol			
10	$134.57	$872.04	$9,298.78
100	$35.57	$129.54	$1,131.28
1000	$25.67	$55.29	$314.53
10000	$24.68	$47.86	$232.86

To additionally test the flexibility, we designed a second set of 3,918 probes to evaluate the methylation state 1 kbp upstream and downstream of 120 genomic regions previously known and confirmed by BSPP to carry aberrant methylation in induced pluripotent stem cells15. We acquired the oligonucleotides from a second vendor (LC Sciences). Even with shorter capturing sequences (40 bp total for capturing arms rather than 50 bp on average, **Figure 2-7a-b**) and a 100-fold smaller target size, an average of 56% of mappable bases were on-target, equivalent to an enrichment factor of ~6,500. With the data from three cell lines (H1 ESCs, PGP1 fibroblasts and PGP1-iPSCs) we identified regions of aberrant methylation in induced pluripotent stem cells and demonstrated that aberrant methylation continues further upstream and downstream than observed previously. This analysis demonstrated that a focused probe set can be

used to validate specific regions of interest identified in global scanning using either our genome-wide probe set or other methods.

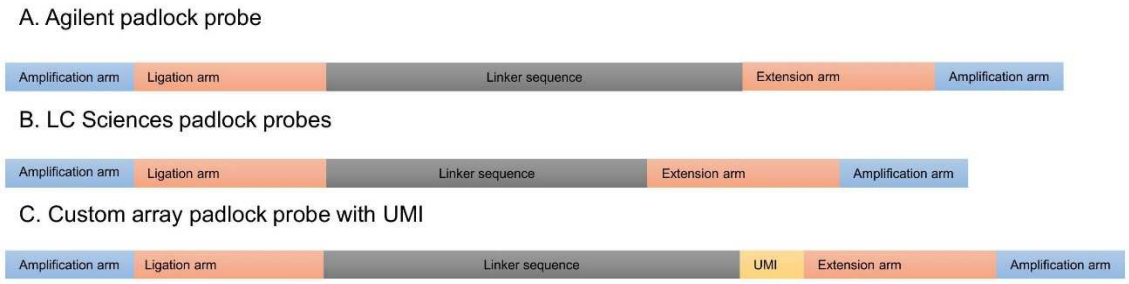


Figure 2-7. Schematic for padlock probes

Genomic coverage of BSPP compared with RRBS

In **Table 2-2**, we estimated that the cost per sample of generating sequencing libraries using the RRBS and BSPP protocols are very similar, with RRBS being slightly less ($28 versus $38). To identify areas where each protocol is advantageous over the other, we applied RRBS and BSPP to ten primary human bone marrow samples and performed an average of 2.3 Gbp and 2.4 Gbp of sequencing per sample for BSPP and RRBS libraries respectively. We used a cutoff of 5X depth of coverage to identify CpGs with coverage across all ten samples. For RRBS and BSPP, we obtained a total of 924,381 and 246,278 CpG sites respectively. Since it is able to cover 3.75 times as many CpG sites as BSPP, the cost per CpG coverage is much better for RRBS. We also found very little overlap between the two assays, only 30,493 CpG sites or 12.4% and 3.3% of BSPP and RRBS CpG sites were overlapping between the two protocols. Even though BSPP have less CpG coverage, we were able to identified ~600 differentially methylated CpG sites specific to human induced pluripotent stem cells[42] which were generally not covered (only 1 CpG was covered) and therefore not discovered by another study that applied the RRBS protocol to identify differentially methylated region specific to human induced pluripotent stem cells[69]. When compared

side by side, the CpG sites from BSPP is more enriched for functional genomic elements than RRBS (**Figure 2-8**).

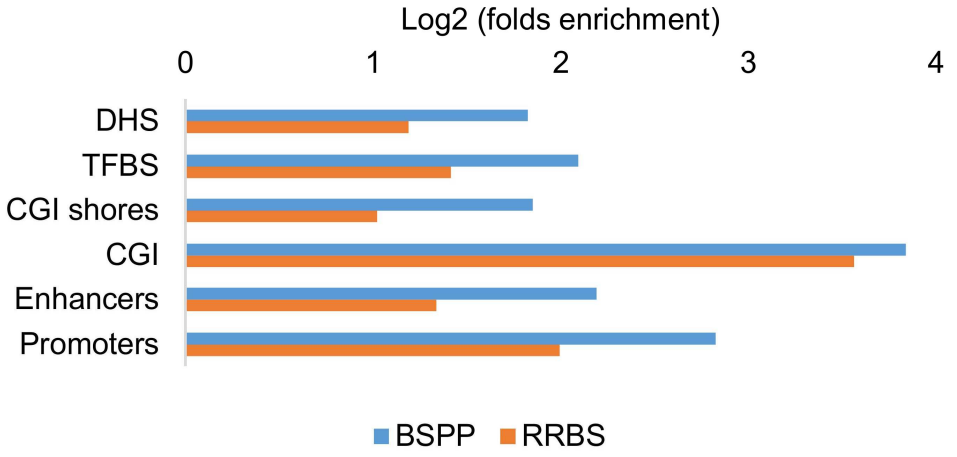


Figure 2-8. Functional genomic region enrichment analysis of captured CpG sites Enrichment factor is computed for annotated genomic region sets. DHS denotes D*nasel* hypersensitive regions identified by ENCODE. TFBS denotes transcription factor binding sites identified by ENCODE. CGI shores are CpG island shores. CGI are CpG islands.

Unique molecular identifiers

Appending a unique molecular identifier (UMI) randomly to padlock probes during oligonucleotide synthesis is a strategy that allows for identification of clonal reads by matching UMIs (**Figure 2-7c**). We performed capture using probes that have been synthesized with ten random bases and quantified the methylation levels for each sample separately for Watson and Crick strands. We found significant improvement in the correlation between the strands when clonal reads were removed with the help of UMIs ($n=8$, $p < 8.1473E-06$, Student's T-Test). With the combinatorial possibilities of ten random bases, we can sample up to 100,000 cells with less than 0.005 probability that two or more different molecules will collide with the same identifier. Absolute counts also enabled us to estimate the true capture efficiencies of individual padlock probes (**Table 2-4**).

Table 2-4. Absolute capture efficiencies

# Molecules captured	% Capable probes
At least 1 from 60 molecules	4%
At least 1 from 600 molecules	39%
At least 1 from 5000 molecules	95%

Application to hydroxymethylation quantification

We performed 5hmC capture with our DMR220K padlock probes in H1 human embryonic stem cells to demonstrate the utility of padlock probes to capture Tet-assisted bisulfite converted DNA. Furthermore, we prepared oxBS-seq libraries from OVCAR3 cell line genomic DNA and successfully performed capture with a modified capture protocol (increased probes to target ratio to 1000:1) using a 80-bp gapsize probeset with ~ 12K padlock probes. While the DMR220K probes were not optimally designed for 5hmC quantification, we were still able to get nearly 4 times the number of 5hmC per gigabase pair sequenced (Gbps) compared with TAB-seq.

Conclusion

In summary, the second-generation BSPP method is scalable, cost effective, and provides absolute quantification of CpG and non-CpG methylation across a large number of highly informative genomic regions. It can be applied to methylation studies on population cohorts with sample sizes up to thousands, and greatly aids in identifying the effects of DNA methylation on human diseases. We also developed a bioinformatics pipeline that can minimize many possible sources of errors in analysis, such as reads trimming to remove sequencing bias, adaptor trimming to remove erroneous methylation calls, high specificity reads alignment approach, and samples identification with SNP

calling. Together these two tools can further advance DNA methylation biomarker development.

Methods

Bisulfite padlock probe production (Oligonucleotides from Agilent)

Libraries of oligonucleotides (~150 nt) were synthesized by ink-jet printing on programmable microarrays (Agilent Technologies) and released to form a combined library of 330,000 oligonucleotides. The oligonucleotides were amplified by PCR in 96 reactions (100 µl each) with 0.02 nM template oligonucleotide, 400 nM each of pAP1V61U primer and AP2V6 primer (Supplementary Table 2), and 50 µL of KAPA SYBG fast Universal 2x qPCR Master Mix (Kapabiosystems) at 95 ºC for 30 s, 15-16 cycles of 95 ºC for 3 s; 55 ºC for 30 s; and 60 ºC for 20 s, and 60 ºC for 2 min. The amplified amplicons were purified by ethanol precipitation and re-purified with Qiaquick PCR purification columns (Qiagen). Approximately 20 ug of the purified amplicons were digested with 50 units of Lambda Exonuclease (5 U/µL; New England Biolabs (NEB)) at 37 ºC for 1 h in lambda exonuclease reaction buffer.

The resulting single-strand amplicons were purified with Qiaquick PCR purification column. Approximately 5-8 µg of single strand amplicons were subsequently digested with 5 units USER (1 U/µl, NEB) at 37 ºC for 1 h. The digested DNAs were annealed to 5.88 µM RE-DpnII-V6 guide oligo (Supplementary Table 2) and denatured at 94 ºC for 2 min decreased the temperature to 37 ºC and incubated at 37 ºC for 3 min. The mixture was digested with 50 units *Dpn*II (10U/µl, NEB) in NEBuffer DpnII at 37 ºC for 2 h. Then the mixture was further digested with 5 units USER at 37 ºC for 2 h followed by enzyme inactivation at 75 ºC for 20 min. The USER/*Dpn*II digested DNAs

were purified with Qiaquick PCR purification column. The single-strand 102 nucleotide probes were purified with 6% denaturing PAGE (6% TB-urea 2D gel; Invitrogen).

Bisulfite padlock probe production (Oligonucleotides from LC Sciences)

The oligonucleotides (100nt) were synthesized using a programmable microfluidic microarray platform (LC Sciences) and released to form a mix of 4,000 oligonucleotides. The oligonucleotides were amplified by two-step PCR in a 200 µL reaction with 1nM template oligonucleotides, 400 nM each of eMIP_CA1_F primer and eMIP_CA1_R primer (Appendix, **Table A-1**), and 100 µL of KAPA SYBR fast Universal qPCR Master Mix at 95 ºC for 30 s, 5 cycles of 95 ºC for 5 s; 52 ºC for 1 min; and 72 ºC for 30 s, 10-12 cycles of 95 ºC for 5 s; 60 ºC for 30 s; and 72 ºC for 30sec, and 72 ºC for 2 min. The resultant amplicons were purified with Qiaquick PCR purification columns and re-amplified by PCR in 32 reactions (100 µL each) with 0.02nM first round amplicons, 400nM each of eMIP_CA1_F primer and eMIP_CA1_R primer, and 50 µL of KAPA SYBR fast Universal qPCR Master Mix at 95 ºC for 30 s, 13-15 cycles of 95 ºC for 5 s; 60 ºC for 30 sec; and 72 ºC for 30 s, and 72 ºC for 2 min.

The resultant amplicons were purified by ethanol precipitation and re-purified with Qiaquick PCR purification columns as described above. Approximately 4 µg of the purified amplicons were digested with 100 units of Nt.AlwI (100 U/ µL, NEB) at 37 ºC for 1 h in NEBuffer 2. The enzyme was heat inactivated at 80 ºC for 20 min. The digested amplicons were then incubated with 100 units of Nb.BrsDI (10 U/ µL, NEB) at 65 ºC for 1 h. The nicked DNA was purified by Qiaquick PCR purification column. The probe molecules (with size of approximately 70 bases) were purified by 6% denaturing PAGE (6% TB-urea 2D gel).

Sample preparation and capture

Genomic DNA was extracted using the AllPrep DNA/RNA Mini kit (Qiagen) and bisulfite converted with the EZ-96 DNA methylation Gold kit (Zymoresearch) in 96-well plate. Normalized amount of padlock probes, 200 ng of bisulfite converted gDNA, and 4.2 nM oligo suppressor were mixed in 25 µL 1x Ampligase Buffer (Epicentre) in 96-well plate, denatured at 95 ºC for 10 min, gradually lowered the temperature at 0.02 ºC/s to 55 ºC in a thermocycler, and hybridized at 55 ºC for 20 h. 2.5 µL of SLN mix (100 µM dNTP, 2 U/µL AmpliTaq Stoffel Fragment (ABI) and 0.5 U/µL Ampligase (Epicentre) in 1X Ampligase buffer) was added to the reaction for gap-filling reaction. For circularization, the reactions were incubated at 55 ºC for 20 h, followed by enzyme inactivation at 94 ºC for 2 min. To digest linear DNA after circularization, 2 µL of exonuclease mix (10 U/µL exonuclease I and 100 U/ µL exonuclease III, USB) was added to the reactions, and the reactions were incubated at 37 ºC for 2 h then inactivated at 94 ºC for 2 min.

TAB-treated genomic DNA were generously provided by G. Hon. We performed bisulfite conversion and capture using DMR220K padlock probes as specified above.

Generation of oxBS-seq libraries for capture

One microgram of genome DNA from OVCAR3 cell line was sheared using Covaris system to 150 bp average fragment size according to Covaris protocols. The sheared DNA was concentrated using 1.8X volumes Agencourt AMPure beads and eluted with 20 µL volume then end-repaired in a reaction with 1X End-it buffer, 1 mM dNTP, 1 mM ATP, and 1 µL of End-it enzyme mix (Epicentre). The reaction was vortexed gently then spun down and incubated at RT for 45 min. Next, the reaction was purified with 1.8X AMPure beads, allowing 10 minutes to bind because fragment sizes are small. DNA were eluted in 15 µL nuclease free water but beads were carried on to

the next step. The A-tailing reaction was performed with 500 µM dATP (New England Biolabs), 1X Tango buffer (Thermo Fisher), 1 µL of Klenow, exo- (Thermo Fisher), and 1 µL of in-house generated controls DNA mixture. The reaction was incubated at 30 °C for 20 min, 37 °C for 20 min, and 75 °C for 10 min. The dA-tailed DNA was adapted immediately in 25 µL total volume of 1X Tango buffer, 1 µL of high capacity T4 DNA ligase (Thermo Fisher), 20 µM of methylated TruSeq adaptors (Illumina), and 500 µM ATP (New England Biolabs). The reaction was carried out overnight at 16 °C. The adaptor-ligated DNA were purified with AMPure beads using 1.8X but this time washing was done using 80% acetilenitrile 4 times while allowing 1 minute during each wash instead of 30 seconds. Next, the adapted DNA were denatured by adding 22.75 µL with 1.25 µL of 1 M NaOH. The reaction was carried out at 37 °C for 30 min with gentle shaking and then immediately transferred to ice. Next, 2 µL of 20 mM $KRuO_4$ (Alpha Aesar) was added to begin oxidation. The reaction was carried out at 40 °C for 30 min with gentle shaking and then immediately transferred to ice. Next, bisulfite conversion was performed using the Epitect Bisulfite Plus Kit (Qiagen) following protocols and then eluted with 20 µL nuclease-free water. Conversion efficiencies were evaluated using 0.5 µL of eluted DNA. After successful conversion of 5hmC was confirmed, we carried out amplification of oxBS-seq library using compatible primers and KAPA Hifi Uracils+ 2X master mix (KAPA Biosystems). The reaction was carried out using the following protocol on the thermocycler: 98 °C for 45 s, then 18 cycles of 98 °C for 15 s, 60 °C for 30s, and 72 °C for 1 min, followed by 72 °C for 2 min.

Capture circles amplification (Agilent Oligonucleotides)

10 µL circularized DNA was amplified and barcoded in 100 µL reactions with 400 nM each of AmpF6.3Sol primer (**Table 2-5**) and AmpR6.3 indexing primer (**Table 2-5**),

0.4x SYBR Green I (Invitrogen), and 50 µL Phusion High-Fidelity 2x Master Mix (NEB) at 98 ºC for 30 s, 5 cycles of 98 ºC for 10 s; 58 ºC for 20 s; and 72 ºC for 20 s, 9-12 cycles of 98 ºC for 10 s; and 72 ºC for 20 s, and 72 ºC for 3 min.

Capture circles amplification (LC Sciences Oligonucleotides)

10 µL circularized DNA was amplified in a 100 µL reaction with 200nM each of CP-2-FA primer and CP-2-RA primer (**Table 2-5**) and 50 µL KAPA SYBR fast Universal qPCR Master Mix at 98 ºC for 30 s, 5 cycles of 98 ºC for 10 s; 52 ºC for 30 s; and 72 ºC for 30 s, 15 cycles of 98 ºC for 10 s; 60 ºC for 30 s; and 72 ºC for 30 s, and 72 ºC for 3 min. The resultant amplicons with the corresponding expected size of approximately 260 bp were purified with 6% PAGE (6% 5-well gel, Invitrogen) and resuspended in 12 µL of TE buffer. 30% of the gel-purified amplicons were re-amplified and barcoded in a 100 µL reaction with 200nM each of two different sets of primers to enable SE sequencing for both ends of the amplicons (CP-2-FA.IndSol primer and CP-2-RA.Sol primer or Switch.CP-2-FA and Switch.CP-2-RA.IndSol) and 50 µL KAPA SYBR fast Universal qPCR Master Mix at 98 ºC for 30 s, 4 cycles of 98 ºC for 10 s; 54 ºC for 30 s; and 72 ºC for 30 s, and 72 ºC for 3 min.

Generation of RRBS sequencing libraries

To generate RRBS sequencing libraries, 100 ng of gDNA were digested with 20 U of *Msp*I (Thermoscientific) in 1X Tango buffer (Thermoscientific) and 1 ng of unmethylated lambda DNA (Promega) in order to assess for bisulfite conversion rate in 30µL total volume for 3 h at 37 ºC and heat inactivated at 65 ºC for 20 min. Next, 5U of Klenow fragment, exo- (Thermoscientifc) and a mixture of dATP, dGTP, and dCTP (New England Biolabs) were added to *Msp*I-digested DNAs for a final concentration of 1 mM, 0.1 mM, and 0.1 mM for dATP, dGTP, and dCTP, respectively in 32 µL for end-repair

and dA-tailing. The mixture was mixed and incubated at 30 ºC for 20 min, 37 ºC for 20 min, and heat inactivated at 75 ºC for 10 min. dA-tailed DNA was purified with 2X volume of Agencourt AMPure XP beads (Beckman Coulter) and resuspended dA-tailed DNA with 20 µL nuclease-free water without discarding the magnetic beads. dA-tailed DNAs were then ligated to methylated adaptors in 30 µL total volume containing 30 U of T4 DNA ligase, HC (Thermoscientific), 1X Ligation buffer (Thermoscientific), and 500 nM individual TruSeq multiplexing methylated adaptors (Illumina). The ligation mixture was mixed well and incubated at 16 ºC for 20 h, heat inactivated at 65 ºC for 20 min, purified by adding 60 µL of PEG 8000/5M NaCl buffer (Teknova) to adaptor ligated DNA and bead mixture, and eluted in 20 µL of nuclease-free water. Next, the adaptor ligated DNA were bisulfite converted using the MethylCode Bisulfite Conversion kit (Life Technologies) following manufacturer's protocol and eluted in 35 µL of Elution buffer (Life Technologies). Bisulfite treated DNAs were amplified using 5 U of PfuTurboCX (Agilent Technologies) and 300 nM each of TruS_F and TruS_R primers for 14 cycles in 100 µL total volume. PCR products were purified with an equal volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted with 50 µL of 10mM Tris-HCl, pH8.5, pooled in equimolar ratios, and size selected using 6% TBE gels for 150-400 bp. The concentration of sequencing libraries was quantified by qPCR using KAPA Library Quantification kit (KAPA Biosystems). Libraries were sequenced on Illumina HiSeq2500 in RapidRun mode for PE 125 cycles.

Spike-in controls

We prepared separate 50 µL PCR reactions with ZymoTaq polymerase, 200 nM of each forward and reverse primer, 1 ng of Lambda DNA template, and 2.5 mM of each dNTPs. For 5hmC control, we used Lambda_hmC primers with d5hmCTPs instead of

dCTPs. For 5mC control, we used Lambda_mC primers with d5mCTPs instead of dCTPs. For C control, we used Lambda_C primers with dCTPs. Appendix **Table A-2** shows all primer sequences. The cycling conditions were as follows: 95 ºC for 10 min, 30 cycles of 95 ºC for 30 s, 55 ºC for 30 s, then 72 ºC for 1 min, and finally 72 ºC for 7 min. The PCR products were then purified with one QiaQuick column each and quantified with the Nanodrop spectrophotometer. Next, we size-selected the amplicons from a 6% TBE polyacrylamide gel to remove the remaining Lambda DNA template. We performed a 2nd round of amplification using the same PCR conditions as previously but with 0.2 nM size-selected PCR products as templates instead of 1 ng of Lambda DNA.

Conversion efficiencies assessment

The following procedures were used to assess the conversion efficiencies of control DNA post oxidative and or post bisulfite treatments. For each control, we used 0.5 µL of converted DNA as input from a total of 20 µL elution volume as template in a 10 µL PCR reaction with 5 µL of KAPA HiFi Uracil master mix, 3.0 µL water, and 1.5 µM of each primer. For C controls, we used the bisLambda_C primers, and for 5mC controls, we used the bisLambda_mC_F primer with the Lambda_mC_R primer. For assessing 5hmC controls post bisulfite only, we used the bisLambda_hmC_F primer with the Lambda_hmC_R primer. For assessing 5hmC controls post-oxidative bisulfite, we used the bisLambda_hmC primers. The cycling condition were as follows: 98 ºC for 45 s, 25 cycles of 98 ºC for 15 s, 55 ºC for 30 s, and 72 ºC for 30 s, and 72C for 1 min.

For assessment via enzymatic digestion, we incubated 1 µL of PCR reaction for each control in a 10 µL reaction with 10 units of *Sph*I and 1X of NEB buffer 2.1 for 3 hours at 37 ºC followed by 20 min at 65 ºC. We ran all 10 µL of digestion reaction on a PAGE gel to analyze the digestion results.

BSPP read mapping and data analysis (v1.0)

Bisulfite converted data for the test capture experiments was processed using a previous version of *BisReadMapper*. Reference genome is computationally converted by changing all C's to T's on Watson and Crick strands separately. FASTQ reads are encoded by 1) predicting the mapping orientation, 2) converting all predicted forward mapping reads by changing all C's to T's and converting all predicted reverse complementary mapping reads by changing all G's to A's, the original reads are maintained. The bisulfite reads are then mapped to the converted reference separately using SOAP2Align (http://soap.genomics.org.cn/soapaligner.html) with the parameters r=0, v=2 (one mismatch per 40bp sequenced), Paired-End: m=0, x = 400. Alignment files are then combined, and one alignment per read was selected. Original C calls were placed back into the alignment information. Alignments are then converted to pileup format using SamTools (http://samtools.sourceforge.net/). Raw SNPs and methylation frequency files were computed from pileup counts.

BSPP read mapping and data analysis (v1.4)

We first trimmed all PE or SE fastq files using trim-galore version 0.3.3 to remove low quality bases and capture sequence positions. Next, the reads were encoded to map to a three-letter genome via conversion of all C to T or G to A if the read appears to be from the reverse complement strand. Then the reads were mapped using BWA mem version 0.7.5a, with the options "-B2 -c1000" to both the Watson and Crick converted genomes. The alignments with mapping quality scores of less than 5 were discarded and only reads with a higher best mapping quality score in either Watson or Crick were kept. Finally, the encoded read sequences were replaced by the original read sequences in the final BAM files. Overlapping pair end reads were clipped with bamUtils clipOverlap

function. Alignments are then converted to pileup format using SamTools (http://samtools.sourceforge.net/). Raw SNPs and methylation frequency files were computed from pileup counts.

Correlation of methylation levels between two samples

To check if methylation levels were similar between two samples, the Pearson's correlation was calculated on all CpG sites characterizable in both. First, a list of CpG sites with read depth of at least 10 in both samples was generated. The methylation frequencies at these sites were obtained from BisReadMapper output, and input into the statistical package R. Finally, Pearson's correlation for the two samples was computed using the cor() function.

Analysis of differential methylation

To identify sites showing a change in methylation between two samples, a list of CpG sites with read depth of at least 10 in both samples was generated. From the BisReadMapper output, the raw read counts showing methylation and lack of methylation were assembled for each line. Using these counts, a Fisher-Exact Test with Benjamini-Hochberg Multiple Testing Correction (FDR=0.01) was carried out on each CpG site. This resulted in a set of differentially methylated sites between the two lines; at each of these sites, the methylation levels were statistically significantly different. Technical replicates did not show any differential methylation, while different cell types showed a large degree (~33%).

Enrichment analysis of methylation haplotype blocks for known functional elements

Genomic regions with same number and fragment length distribution were randomly sampled within the mappable regions (regions with minimum 10X coverage in

WGBS dataset), and repeated 1,000 times. Statistical significance was estimated based on the number of times an equivalent or higher number of overlapping regions were found. Fold changes (enrichment factors) were calculated as the ratios of observation over random expectation. Enhancer definition was based on Andersson et al.[70], and promoter regions were based on the definition by Thurman et al.[71]. All the genomic coordinates were based on GRCh37/hg19.

Acknowledgements

We thank T. Tanaka and R. Bejar for providing the primary human bone marrow genomic DNA and generating the sequencing data for comparison iof BSPP with RRBS. We also thank A. Feinberg and J. Stamatoyannopoulos for providing informative genomic targets and E. LeProust (Agilent Technology) for long-oligonucleotide synthesis. This work is funded by grants from National Institute of Health (R01 DA025779; R01 GM097253) to Kun Zhang.

Chapter 2 contains material as it appears in: Dinh Diep*, Nongluk Plongthongkum*, Athurva Gore*, Ho-Lim Fung, Robert Shoemaker, Kun Zhang. “Library-free Methylation Sequencing with Bisulfite Padlock Probes.” Nature Methods. 2012 February 5; 9(3): 270-272. doi: 10.1038/nmeth.1871. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3: A generalized method for the identification of differentially methylated regions on shallow WGBS data

Introduction

A major goal of many epigenome mapping studies is to identify regions of dynamic or differential DNA methylation (DMRs). These regions are important for the mapping of methylation patterns, epigenome-wide association studies, and for the identification of epigenetic regulatory events. To identify DMRs from whole genome bisulfite sequencing (WGBS) datasets required either utilized high depth of coverage datasets (30-60X) or a stringent DMR finding criteria for low depth of coverage datasets[43,49,72–74]. Reducing the coverage requirement for DMR finding has the advantage of enabling higher sample sizes for the same cost that can lead to more robust DMRs. Additionally, studies on rare cells or samples that are difficult to attain often results in low coverage datasets and these studies would benefit from a DMR finding method with a low depth of coverage requirement. Finally, 5-hydroxymethylcytosines modifications which predominantly occur at low levels would also benefit from a low coverage DMR finding method that leverages the power of nearby CpG sites to identify differentially hydroxymethylated regions (DHMRs).

There are currently a few DMR finding approaches that consider low depth of coverage datasets and they have mainly focused on pairwise comparison between two types of samples[50,75]. Pairwise comparisons of samples when performed on datasets with multiple groups of samples or undefined samples have exponential compute costs and multiple testing concerns. To overcome the limitations of current approaches, *cgDMR-miner* was developed to identify the most DMRs from generalized datasets, including samples that have not been categorized, without performing all pairwise comparison.

In this chapter, we developed *cgDMR-miner*, a generalized method for the identification of differentially methylated CpG regions (DMRs) on shallow whole genome

bisulfite sequencing (WGBS) datasets with more than two sample groups. The datasets can either be categorized or uncategorized because *cgDMR-miner* employs a variability score to identify the most variably methylated regions. As a key novel feature, this variability score is able to identify differential methylation patterns without being dependent on the methylation levels at individual locus and is the main novelty of this work. This method handles data with low depth of coverage at a higher sensitivity than previous methods, and can be applied to whole genome TET-assisted bisulfite sequencing (TAB-seq) datasets to identify regions of differentially 5-hydroxy-methylated CpG sites.

Results

Comparison with current approaches

The key to *cgDMR-miner* is the JSD metric. In comparison with other scores for quantifying methylation variabilities, JSD was able to discern larger absolute differences from smaller absolute differences while also being unbiased towards hypo- or hyper-methylation (**Figure 3-1**). The second best metric was the root mean square error (RMSE) that was also unbiased towards hypo- or hyper-methylation but an absolute difference of 0.1 versus 1.0 reduced the score linearly by ten times while for JSD the reduction was twenty-four times. Unlike RMSE, JSD is more robust against small changes, and this property makes it useful for segmentation.

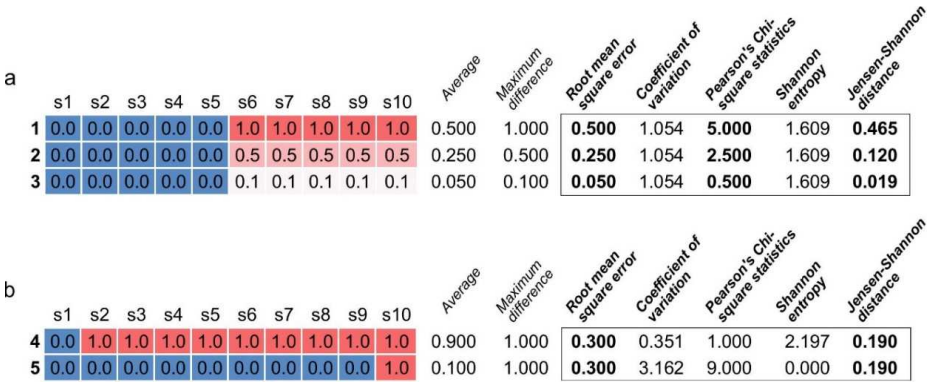


Figure 3-1. Quantifying methylation variabilities

Methylation variabilities are quantified for artificial examples of ten samples (s1 to s10) across five CpG sites (#1-5). In (a), the acceptable metrics are bolded as they exhibit decrease in methylation variabilities for #1 to #3. Furthermore, Jensen-Shannon distance metric is best able to distinguish 0.1 difference from 1.0 difference, since it has 25 folds difference between #1 and #3. In (b), the acceptable metrics should give the same methylation variabilities for examples #4 and #5.

In tests to verify the performance of *cgDMR-miner* at 5X coverage, we found that it consistently outperforms Hon et al. 2013[72] and *MethylPy*[49] from 0.1, 0.2, to 0.4 DMRs difference (**Figure 3-2**).

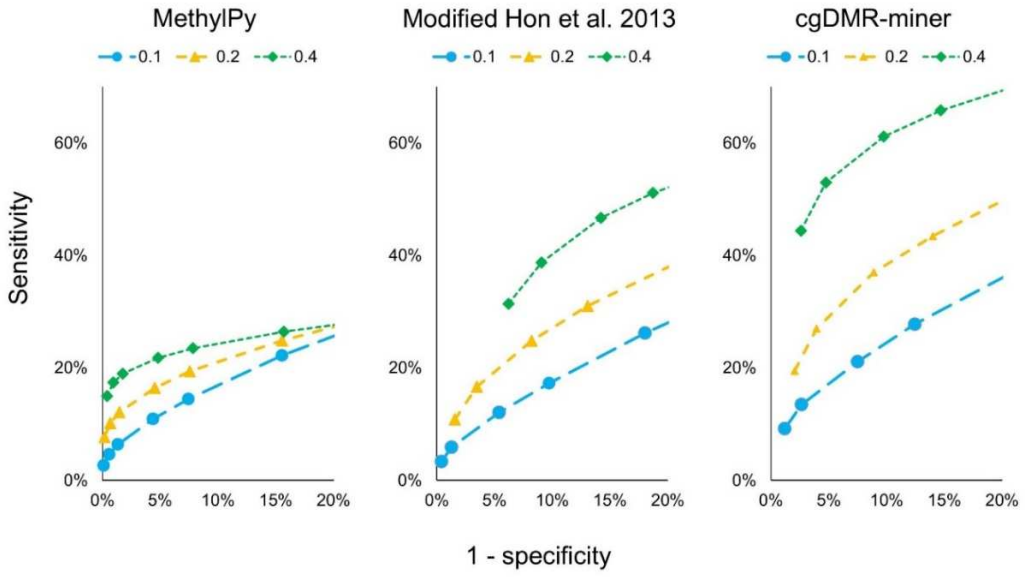


Figure 3-2. Receiver operating characteristics for DMRs with 0.1, 0.2, and 0.4 methylation differences

Receiver operating characteristics curves for *cgDMR-miner*, our implementation of the Hon et al. 2013 method, and MethylPy for simulated datasets with 5X depth of coverage and DMR methylation differences of 0.10, 0.20, and 0.40.

When comparing the DMR CpGs identified by the five different methods at 20X depth of coverage, we found that *cgDMR-miner* was most similar to *MethylPy* (which uses RMSE test) and Hon et al. 2013 (which uses Pearson's Chi-square goodness of fit test) and least similar to Ziller et al. 2013 (**Figure 3-3**).

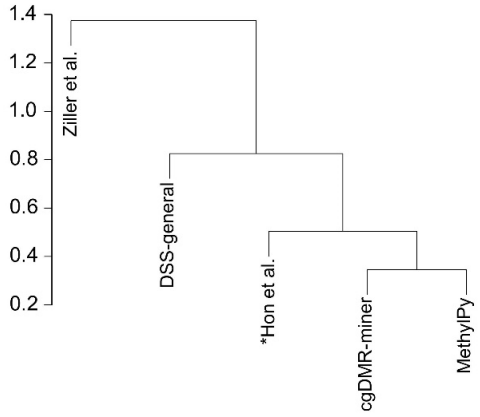


Figure 3-3. Similarity of five methods using DMRs from simulated 20X depth of coverage
Five methods at simulated 20-fold average depth of coverage are compared using the DMRs called by each. Distances are 1 – Pearson's correlation and is calculated on a binary matrix with "1" indicating that a CpG was identified as DMR or "0" indicating that it was not. Only CpGs that have been called DMR in a method are considered. *cgDMR-miner* is most similar to MethylPy and Hon et al. 2013. Ziller et al. 2013 is the most different, probably due to a higher number of false positives. DSS-general is in-between Ziller et al. 2013 and the other methods.

Method overview

cgDMR-miner analyzes multiple chromosomes in parallel or sequentially when processing power is limited, and is comprised of the following analysis steps (**Figure 3-4a**). First, *cgDMR-miner* applies the BSmooth function from the R package *bsseq* to perform local linear smoothing on each sample. The smoothing parameters necessary for BSmooth are determined by 10-folds cross-validation on one segment of the chromosome for each sample. After smoothing, *cgDMR-miner* calculates a cross-sample variability score for each CpG site based on the Jensen-Shannon distance (JSD) (see Methods). This metric is similar to the one described by Ziller et al. *Nature* 2013, except that the distance is calculated against a uniform distribution. Next, segmentation is

performed on the vectors of variability scores using a 5-states Hidden Markov Model. For non-WGBS data, circular binary segmentation[76] is also implemented in *cgDMR-miner*. Finally, a test for homogeneity of proportions is performed on the methylated read counts of each segment. We applied a generalized C(α) binomial goodness of fit test statistics derived by Tarone R. E. 1979 with adaptive permutation to estimate the p-value. This method by Tarone 1979 was shown to be applicable for testing overdispersion in heterogeneous data[77]. Although this test is optimal for small sample sizes (where n is the total read depth), it will generate significant p-values for very large sample sizes with small differences. Thus, we apply a filter by effect size to identify DMRs.

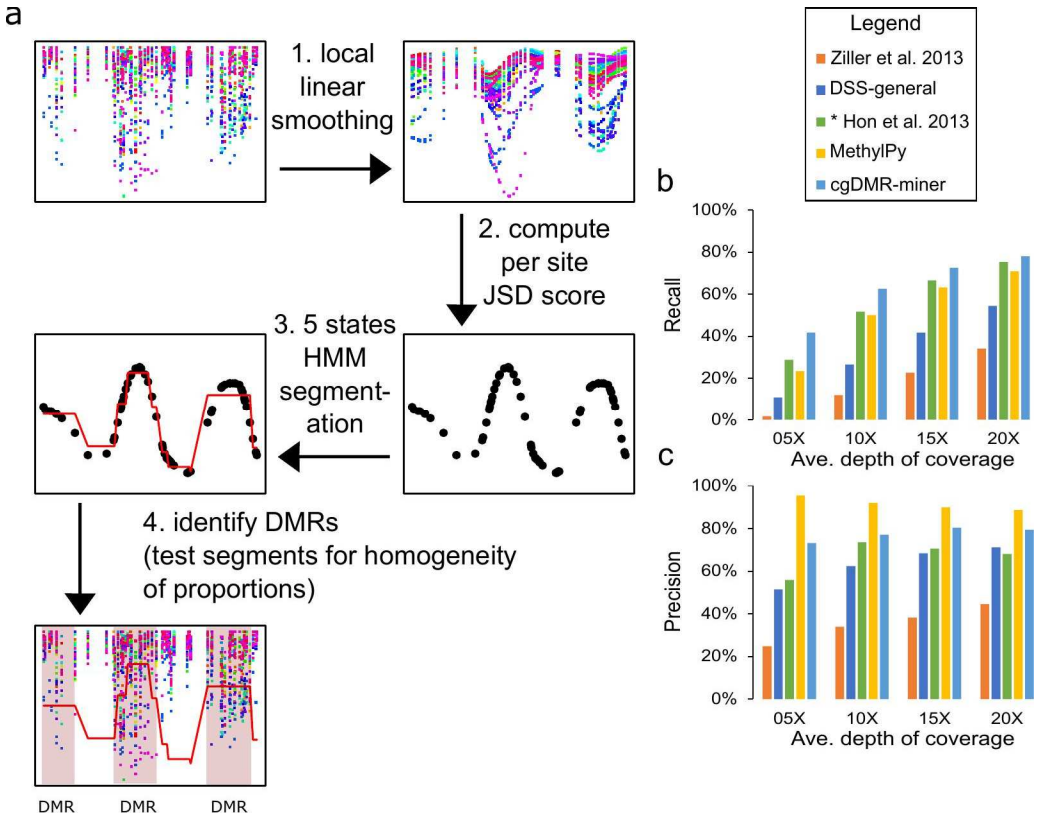


Figure 3-4. Overview of method and performance
Overview of *cgDMR-miner* method and performance comparison. (a) Schematic of *cgDMR-miner*. (b) Recall comparison. (c) Precision comparison.

Usage

To run *cgDMR-miner*, the user launches a Perl script from command-line. Smoothing can require up to 20 GB of random accessed memory (RAM) and will take approximately 25 minutes per sample for human chromosome 1, while smaller chromosomes will take less time. After all the samples and chromosomes have been smoothed, DMR identification on chromosome 1 among 36 samples requires only up to 8 GB of RAM and 22 hours of compute time on a single CPU core. The final outputs include (1) a matrix of the average methylation level of all methylation segments, (2) an information file for the DMRs with: overall average methylation level, total depth of coverage, test statistic, empirical p-value, maximum average methylation levels between two samples, and the standard deviation of the average methylation levels.

Benchmarking

We benchmarked *cgDMR-miner* against previously published methods: (Hon et al. *Nature Genetics* 2013, Ziller et al. *Nature* 2013, Schultz et al. *Nature* 2015 – *MethylPy*, Park et al. *Bioinformatics* 2016 – *DSS-general*[44]). To compare their performances, we generated a simulated dataset for chromosome 19 with 15,498 randomly placed DMRs between 1 to 2,765 bp in lengths with 36 uncategorized samples. The overall number of DMR CpG accounts for 23% of CpGs with an average of 16 CpGs per DMR. We compared *cgDMR-miner* to other methods and found that *cgDMR-miner* outperform in recall rates while being second only to *MethylPy* in precision rates, with 42% recall and 73.3% precision at the 5X depth of coverage (**Figure 3-4b,c**). At the highest simulated depth of coverage of 20X, we found that *cgDMR-miner* was most similar to *MethylPy* and the Hon et al. 2013 approach. Furthermore, when applied to 36 human tissues dataset (Schultz et al. 2015), *cgDMR-miner* identified 5% more DMRs than *MethylPy*. *MethylPy* also required 23 hours with two processing cores to

process chr19 at 30X coverage for 36 samples, while *cgDMR-miner* used just 8.13 hours.

Application to real world data

Next, to assess the capability for DHMR calling, we applied *cgDMR-miner* to a real-world 30X average depth of coverage TAB-seq dataset consisting of matched tumor and normal tissues from two patients diagnosed with clear cell renal carcinoma (ccRCC)[11]. Since the level of 5hmC modification have not been found to be strongly correlation between nearby CpGs, we thought *cgDMR-miner* might perform poorly compared to *MethylPy* because smoothing would dilute some weak signals at the single CpG sites. We applied both methods to the TAB-seq dataset and found that 83% of the DHMRs identified by cgDMR-miner overlap with the DHMRs identified by *MethylPy*. However, *cgDMR-miner* have also identified 78% of the DHMRs identified by *MethylPy*.

Tissue specific DMRs have been shown to have significant enrichment of transcription factor binding sites[72,74], here, we found that both *cgDMR-miner* and MethylPy can recover a higher fractions of DMRs overlapping with transcription factor binding (**Figure 3-5**). While a higher proportion of DMRs by MethylPy overlapped with a transcription factor binding site, *cgDMR-miner* was able to identify more DMRs overlapping with transcription factor binding. For the total number of TFBS overlapping DMRs, *cgDMR-miner* identifies 332,473 (out of 737,084 DMRs called), MethylPy identifies 313,656 (out of 626,418 DMRs called), and Ziller et al. 2013 approach identifies 337,958 (out of 1,198,131 DMRs called).

Chen et al. *Cell Research* 2016 demonstrated that genes body with loss of hydroxymethylation and gain of methylation were significantly downregulated in kidney cancer. Using *cgDMR-miner* to identify DMRs with loss of hydroxymethylation and gain of methylation in kidney cancer, we identified about ten times more DMR CpGs (~2

million DMR CpGs), some of these would be missed even if MethylPy was used (**Figure 3-6**). Re-analysis of RNA-seq data from the study allowed us to verify that the genes overlapping with the new set of DMRs were downregulated in kidney cancer as well (**Figure 3-7**).

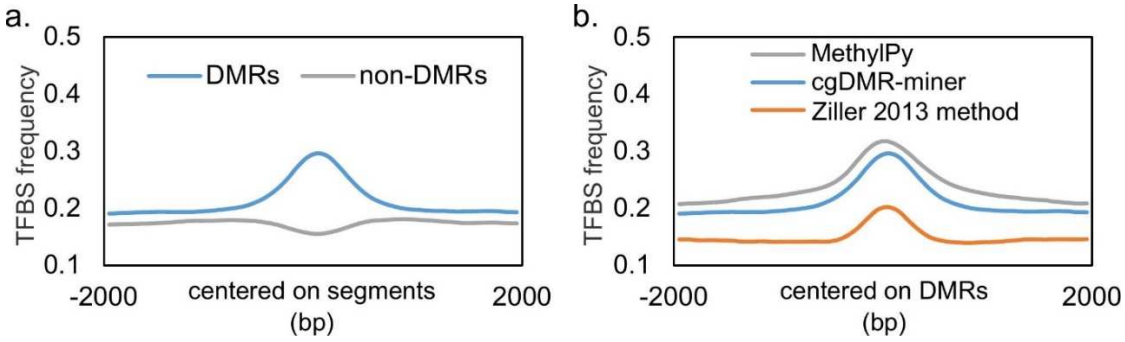


Figure 3-5. Differentially methylated regions overlap with transcription factor binding sites
DMRs identified from 36 human tissues are overlapped with a list of all predicted transcription factor binding sites. In (a), the DMRs versus non-DMRs segments from *cgDMR-miner* are compared. In (b), the DMRs from *cgDMR-miner* are compared against the DMRs from Schultz et al. 2015 using either Ziller et al. 2013 approach or using MethylPy.

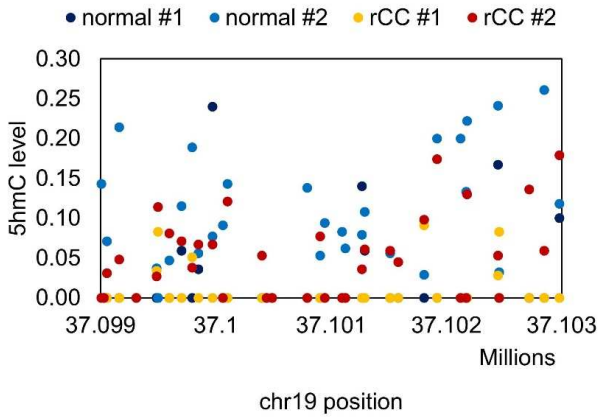


Figure 3-6. Example of DHMR missed by MethylPy
MethylPy missed a DHMR region. This region is found in the gene body of *ZNF382*, which encodes the transcription factor KZNF. KZNF have been associated with playing a critical role as tumor suppressor in multiple carcinomas.

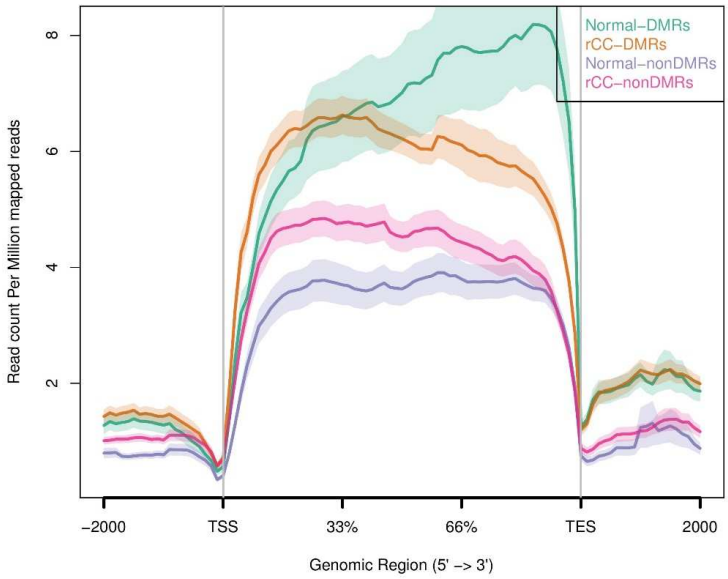


Figure 3-7. Hypo-DHMRs with hyper-DMRs in kidney cancer tissues
RNA-seq data shows expression profile of a ccrCC kidney (rCC) compared to the profile of a normal kidney. The "DMRs" genes (n=8,524) have gene bodies overlapping with 95,402 regions (covering 2,092,132 CpGs) with hypo-hydroxymethylation and hyper-methylation in ccrCC kidney versus normal kidney for two patients. In DMR overlapping genes, ccrCC kidney appears to be downregulated compared to normal kidney while the profiles for other non-DMRs genes has higher expression for ccrCC kidney. Using a site-wise comparison method, Chen et al. identified only 211,519 and 230,341 CpGs for patient 1 and 2 respectively. Thus, smoothing can also be applied to TAB-seq datasets to identify DHMRs.

Conclusion

We demonstrated the usage of *cgDMR-miner* in simulated and real- world data and showed improved sensitivity for DMR finding in low coverage and 5-hydroxymethylation datasets while maintaining a level of precision around 70% that is better than most current methods. With two processing threads, *cgDMR-miner* is almost 300% faster than other methods that perform analysis at individual CpG sites. Most importantly, we demonstrated that smoothing and segmentation improved the sensitivity of DMR finding by combining neighboring sites with similar signals together.

Methods

Data processing and (hydroxy)methylation calling from WGBS and TAB-seq datasets

Published WGBS (SRP000941) and TAB-seq datasets (SRP049710) are from the short reads archive (SRA). BisReadMapper (https://github.com/hdinhdp/bisReadMapper) is used align fastq reads to hg19/GRCh37. Methylation frequency files are generated using the aforementioned mapping and methylation analysis pipeline. Custom in-house scripts are used to re-format the methylation frequency files to the appropriate input files for each DMR calling software used in this study. A full list of published data utilized in this study is in **Table 3-1**.

Data processing and analysis of RNA-seq datasets

Published RNA-seq datasets (SRP049710) are from the short reads archive (SRA). STAR (https://github.com/alexdobin/STAR) is used to align fastq reads to hg38. To map DMRs to genes, Gencode reference version 75 (for GRCh37) was used. The ensemble id for DMR genes were extracted and given as an input list to ngsplot-2.61 (https://github.com/shenlab-sinai/ngsplot) to generate genes expression profile plots.

Quantification of methylation variability

Various metrics for quantifying methylation variabilities can be considered. First, the Pearson's chi-square statistics can be calculated using the average as the expected value. The Shannon entropy is calculated as follows: first the methylation frequency vector is transformed into a probability distribution by dividing each methylation frequency by the sum of all the methylation frequencies. The Jensen-Shannon distance (JSD) is determined using the same transformation of methylation frequencies into a probability distribution. Also, the smaller of either the JSD for the methylated frequencies from uniform or the JSD for the unmethylated frequencies from uniform.

Differential methylation analysis with cgDMR-miner

Smoothing raw methylation frequencies. Each WGBS methylation frequency data is smoothed using local linear smoothing (R package **BSmooth**). Building the smoothing model requires the smoothing parameters h which is the minimum smoothing window size and ns which is the minimum number of site per smoothing window. Cross-validation was performed to identify the optimal smoothing parameters for the data. The parameters for smoothing is determined for each chromosome using a segment with 100,000 CpG sites not more than 50,000 bp apart between adjacent CpGs. 10% of sites is randomly selected as the validation set and remaining 90% as training set. The h parameter is kept constant first at 500 bp while values for the parameter ns is first tested in increment of 2 and from 14 to 50. The lowest value that generates the highest correlation of methylation level with the validation set is chosen. Next, the ns value is kept constant at the chosen value, and the h parameter is tested in increment of 400 from 500 to 2000. The lowest h value with the highest correlation of methylation level with the validation set is chosen. Each chromosome is then smoothed using the chosen parameters.

Computing the Jensen-Shannon distance. The Jensen-Shannon distance (JSD) is calculated for each smoothed methylation vector at individual CpG sites. For each smoothed methylation vector, the missing values were also set equal to the median first. Smoothed methylation frequencies are converted to probability distributions **O** = (o_1,o_2,o_3,…,o_N), **P** = (p_1,p_2,p_3,…,p_N), and **Q** = (q_1,q_2,q_3,…q_N) with elements that are defined by Eq. 1, Eq. 2, and Eq. 3.

Eq. 3-1.
$$o_t = mt / \sum_{t=1}^{N} m_t$$

Eq. 3-2.
$$p_t = (1 - mt) / \sum_{t=1}^{N} (1 - mt)$$

Eq. 3-3. $$q_t = 1/\sum_{t=1}^{N} 1$$

The Jensen-Shannon Distance (D_{JS}) for probability distributions X and Y is defined by Eq. 4. The smaller distance between $D_{JS}(O||Q)$ and $D_{JS}(P||Q)$ is assigned to the JSD score.

Eq. 3-4. $$D_{JS}(X||Y) = \sqrt{\frac{1}{2}D_{KL}(X||M) + \frac{1}{2}D_{KL}(Y||M)}$$

Where M is the average of two probability distributions, and D_{KL} is the Kullback-Leibler divergence which is the measure of difference between probability distributions. M and D_{KL} are from Eq. 5 and Eq. 6.

Eq. 3-5. $$M = \frac{1}{2}(X + Y)$$

Eq. 3-6. $$D_{KL}(A||B) = \sum_i A_i log \frac{A_i}{B_i}$$

Segmentation. The JSD scores along each chromosome are used for genome segmentation. Whole genome data is segmented using a 5-states Hidden Markov Model with a single Gaussian as the emission distribution for each state. The HMM is initialized with equal starting probabilities, a transition matrix which allows only stepwise changes from one state to the next, and starting emission Gaussians with means that are the sorted top five bins with highest frequencies and standard deviations that are simply the entire sample's standard deviation. The R package **hsmm** performs expectation-maximization to estimate the model and then global decoding to determine the hidden state sequence with the Viterbi algorithm. Non-whole genome data can be segmented using circular binary segmentation with the smooth.CNA function of the **DNAcopy** R

package. The default parameters were used in smooth.CNA. Finally, CpG sites with the same hidden state and within 1000 bp of each other are merged to form segments.

Identifying differentially methylated segments. The segments identified are regions that can be assumed to have consistent methylation levels across CpG sites and consistent variabilities across samples. Each segment is then summarized using the total CpG coverage and total methylated CpG counts. To minimize the noise from randomly sampling of bisulfite reads, only regions with a minimum total CpG coverage of 10 in at least 2 samples is considered. The expected methylation count for sample i, m_{ei} , is computed using Eq. 7 where d_i is the total CpG coverage for sample i and e_{mf} is the estimated expected methylation frequency of each segment assuming that methylation frequencies observed fit the binomial model with mean e_{mf}. The test statistic for the null hypothesis that each methylated counts are independent binomial random variables from a binomial model is defined by Eq. 8 where N is the number of samples, and the observed methylation count for sample i is m_{oi}. This test statistic have an asymptotic chi-squared distribution with one degree of freedom (Tarone R. E. 1979). An empirical p-value for each statistics is computed by adaptive permutation where the methylation counts are generated using a binomial distribution with d_i total number of trials and e_{mf} probability of being methylated. Finally, the regions with effect sizes passing threshold and empirical p-value thresholds ($p < 0.01$) are identified as DMRs, thus rejecting the null hypothesis that a single binomial model fit the observed proportions for each sample.

Eq. 3-7
$$m_{ei} = e_{mf} d_i$$

Eq. 3-8
$$\chi_1^2 = \frac{\left\{\sum_i^N \frac{(m_{oi} - m_{ei})^2}{e_{mf}(1-e_{mf})} - \sum_i^N d_i\right\}^2}{2\left\{\sum_i^N d_i(d_i - 1)\right\}}$$

Analysis of simulated datasets using previous methods

Hon et al. 2013. A modified version of the method described by Hon et al. 2013 is implemented using custom scripts. Similar to the published method, a chi-square statistic is generated along the genome using a 3-CpGs sliding window. Next, the chi-square statistics are transformed using the natural log of 10. The 0 values are approximated to 0.0001. Hidden Markov Model segmentation is performed using a 4 states HMM with 3 different emission distributions, one for each hidden state. The emission distributions are modeled as Gaussians. After segmentation, the adjacent windows with the same states are merged together unless they are more than 500 bp apart. Next, the DMRs are identified using the approach described in *cgDMR-miner* with permutation p-value cutoff of < 0.01.

Methylpy. Methylpy from Schultz et al. 2015 is utilized with no modification. The analyses use these parameters as followed: `num_sims` = 3000, `num_sig_tests`=100, `dmr_max_dist`=250, `mc_max_dist`=100, `sig_cutoff`=0.05, `min_cov`=1. The `mc_max_dist` parameter allows for adjacent CpGs within the set bp value to be counted together in low coverage datasets but also decrease the specificity.

Ziller et al. 2013. The analysis based on Ziller et al. 2013 is generated using the scripts from Schultz et al. 2015. The analysis of the simulated datasets has no modification.

DSS-general. One-versus-all analyses are performed using DSS-general with no modification. The DML are identified using the *DMLfit.multiFactor* and *DMLtest.multiFactor* functions. The DMRs are identified using the *callDMR* function with these parameters: `minlen`=1, `minCG`=1, and `p.threshold`=0.05.

Simulated datasets

The simulated datasets are based on chromosome 19 and methylation frequency files are generated for 36 samples. Individual CpG coverages are simulated using random sampling from the Poisson distribution with means that are estimated by the total coverage across 36 real human tissues WGBS datasets. The total coverages were normalized to the equivalence of 5X, 10X, 15X, or 20X average depth of coverages in different simulated datasets. Individual CpG methylated counts are simulated using random sampling from the Binomial distribution with the coverages as the number of trials and the methylation frequency across 36 real human tissues WGBS datasets.

The minimum lengths of differentially methylated regions are simulated by 20,000 random sampling of a normal distribution with mean 500 bp and standard deviation of 500 bp. Each potential DMR is assigned to a CpG randomly then extended to the next CpG until the DMR's expected length is reached but if there is no adjacent CpG to extend the DMR to the expected length, the potential DMR is dropped. The samples are randomly assigned to a DMR, and the number of samples to assign to a DMR is determined by a random sampling of a Poisson distribution with mean of 0.3. The methylated counts for a sample with the DMR are simulated using random sampling from the Binomial distribution with the methylation frequency difference of 0.1, 0.2, or 0.4 for different simulated datasets. The number of hypermethylated or hypomethylated DMRs are estimated from Schultz et al. 2015 DMRs, which is ~ 8% hypermethylated and ~92% hypomethylated.

Recall and precision calculation

Simulated datasets with 5X, 10X, 15X, or 20X average coverages and DMR methylation frequency difference of 0.3 are analyzed using *cgDMR-miner* or alternative approaches. The recall rate is the number of true DMR CpG identified determined as

within a DMR. The precision rate is the number of true DMR CpG identified out of all the CpGs determined as within a DMR.

Receiving operator characteristics curves

The simulated datasets with 5X average coverages and DMR methylation frequency difference of 0.1, 0.2, and 0.4 are analyzed using *cgDMR-miner* or alternative approaches. Different p-value cutoffs from 0.001 to 1 are used to generate the receiving operator characteristics curves at different methylation difference levels for DMRs.

Transcription factor binding sites enrichment analysis

References for transcription factor binding sites (TFBS) are from the ENCODE project. The segments of interest are randomly sampled 100,000 times without replacement. Each segment is extended 2000 bp to the left and to the right to generate a 4000 bp region that is then split into 40 windows of 100 bp in sizes. Then each window is overlap with the set of defined TFBS and the number of bases within the window that is overlapping a TFBS is counted and averaged over each window.

Supplementary Tables

Table 3-1. List of published datasets analyzed

Sample	Method	Source	Tissue	Depth of coverage
STL001BL-01	WGBS	Roadmap Epigenetics Project	Bladder	78
STL001FT-01	WGBS	Roadmap Epigenetics Project	Fat	28
STL001GA-01	WGBS	Roadmap Epigenetics Project	Gastric	30
STL001LG-01	WGBS	Roadmap Epigenetics Project	Lung	28
STL001LV-01	WGBS	Roadmap Epigenetics Project	Heart	64
STL001PO-01	WGBS	Roadmap Epigenetics Project	Muscle	27
STL001RV-01	WGBS	Roadmap Epigenetics Project	Heart	25
STL001SB-01	WGBS	Roadmap Epigenetics Project	Intestine	70
STL001SG-01	WGBS	Roadmap Epigenetics Project	Colon	75
STL001SX-01	WGBS	Roadmap Epigenetics Project	Spleen	37
STL001TH-01	WGBS	Roadmap Epigenetics Project	Thymus	66
STL002AD-01	WGBS	Roadmap Epigenetics Project	Kidney	34
STL002AO-01	WGBS	Roadmap Epigenetics Project	Vessel	25
STL002EG-01	WGBS	Roadmap Epigenetics Project	Esophagus	31
STL002FT-01	WGBS	Roadmap Epigenetics Project	Fat	36
STL002GA-01	WGBS	Roadmap Epigenetics Project	Gastric	27
STL002LG-01	WGBS	Roadmap Epigenetics Project	Lung	75
STL002OV-01	WGBS	Roadmap Epigenetics Project	Ovary	76
STL002PA-01	WGBS	Roadmap Epigenetics Project	Pancreas	30
STL002PO-01	WGBS	Roadmap Epigenetics Project	Muscle	31
STL002SB-01	WGBS	Roadmap Epigenetics Project	Intestine	22
STL002SX-01	WGBS	Roadmap Epigenetics Project	Spleen	34
STL003AD-01	WGBS	Roadmap Epigenetics Project	Kidney	72
STL003AO-01	WGBS	Roadmap Epigenetics Project	Vessel	112
STL003EG-01	WGBS	Roadmap Epigenetics Project	Esophagus	81
STL003FT-01	WGBS	Roadmap Epigenetics Project	Fat	65
STL003GA-01	WGBS	Roadmap Epigenetics Project	Gastric	78
STL003LV-01	WGBS	Roadmap Epigenetics Project	Heart	70
STL003PA-01	WGBS	Roadmap Epigenetics Project	Pancreas	62
STL003PO-01	WGBS	Roadmap Epigenetics Project	Muscle	81
STL003RA-01	WGBS	Roadmap Epigenetics Project	Heart	77
STL003RV-01	WGBS	Roadmap Epigenetics Project	Heart	72
STL003SB-01	WGBS	Roadmap Epigenetics Project	Intestine	28
STL003SG-01	WGBS	Roadmap Epigenetics Project	Colon	65
STL003SX-01	WGBS	Roadmap Epigenetics Project	Spleen	72
STL011LI-01	WGBS	Roadmap Epigenetics Project	Liver	39
P1-T	WGBS	GSE63183	Kidney cancer	22
P1-N	WGBS	GSE63183	Kidney	23
P2-T	WGBS	GSE63183	Kidney cancer	22
P2-N	WGBS	GSE63183	Kidney	20
P1-T	TAB	GSE63183	Kidney cancer	29
P1-N	TAB	GSE63183	Kidney	16
P2-T	TAB	GSE63183	Kidney cancer	33
P2-N	TAB	GSE63183	Kidney	29

Acknowledgements

We thank the members of the Zhang lab and Y. He for insightful discussions.

Chapter 3 contains material from a submitted manuscript: Dinh Diep* and Kun Zhang. "*cgDMR-miner*: generalized method for the identification of differentially methylated regions from low coverage datasets". The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4: Deconvolution of epigenetic heterogeneity in human tissues and plasma DNA by tightly coupled CpG methylation

Introduction

CpG methylation in mammalian genomes is a relatively stable epigenetic modification, which can be transmitted across cell division[5] through the DNA methyltransferase DNMT1 and dynamically either established or removed by the DNMT3A, DNMT3B and ten-eleven translocation proteins (TETs). Due to the processivity of some of these enzymes, physically adjacent CpG sites on the same DNA molecules can share similar methylation status, although discordant CpG methylation has also been observed, especially in cancer cells[78]. The theoretical framework of linkage disequilibrium[79], which was developed to model the coordinated segregration of adjacent genetic variants on human chromosomes among human populations, can be applied to the analysis of CpG co-methylation in cell populations. A number of studies related to the concepts of methylation haplotypes[59], epi-alleles[80], or epi-haplotypes[81] have been reported, although at small numbers of genomic regions or limited numbers of cell and tissue types. Recent data production efforts, especially by large consortia[82–84], have produced a large number of whole-genome, base-resolution bisulfite sequencing data sets for many tissue and cell types. These public data sets, in combination with additional whole-genome bisulfite sequencing (WGBS) data generated in this study, allowed us to perform full-genome characterization of locally coupled CpG methylation across the largest set of human tissue types available to date and to annotate these blocks of co-methylated CpGs as a distinct set of genomic features.

DNA methylation is cell-type specific, and the pattern can be harnessed for analyzing the relative cell composition of heterogeneous samples, such as different white blood cells in whole blood[85], fetal components in maternal circulating cell-free DNA(cfDNA)[31], or circulating tumor DNA (ctDNA) in plasma[31]. Most of these recent efforts rely on the methylation level of individual CpG sites, and they are fundamentally

limited by the technical noise and sensitivity in measuring single-CpG methylation. Recently, Lehmann-Werman *et al.* demonstrated superior sensitivity with multi-CpG haplotypes in detecting tissue-specific signatures in cfDNA[30], although this was based on the sparse genome coverage of Illumina 450k methylation arrays (HM450K). Here we performed an exhaustive search of tissue-specific methylation haplotype blocks (MHBs) across the full genome and proposed a block-level metric, termed methylated haplotype load (MHL), for a systematic discovery of informative markers. By applying our analytical framework and identified markers, we demonstrate accurate determination of tissue origin and prediction of cancer status in clinical plasma samples from patients with lung cancer (LC) or colorectal cancer (CRC) (**Figure 4-1a**).

Results

Identification of methylation haplotype blocks

To investigate the co-methylation status of adjacent CpG sites along single DNA molecules, we extended the concept of genetic linkage disequilibrium[59,79,86] and the r^2 metric to quantify the degree of coupled CpG methylation among different DNA molecules. Methylation status of multiple CpG sites in single- or paired-end Illumina sequencing reads were extracted to form methylation haplotypes, and pairwise 'linkage disequilibrium' of CpG methylation r^2 was calculated from the fractions of different methylation haplotypes.

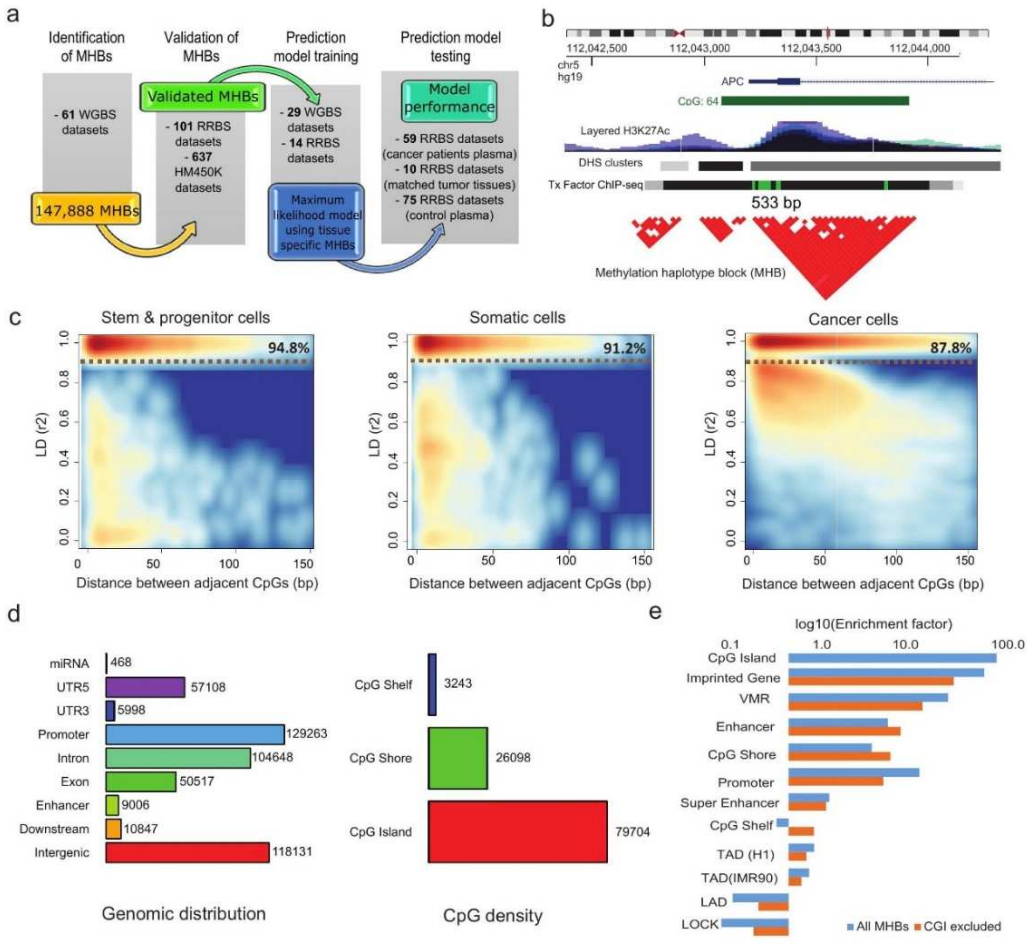


Figure 4-1. Identification and characterization of human methylation haplotype blocks (MHBs). (a) Schematic overview of data generation and analysis. (b) An example of an MHB at the promoter of the gene APC. Tx, transcription; DHS, DNA hypersensitive sites. (c) Smooth scatter plots of methylation linkage disequilibrium within MHBs in stem and progenitor cells (left), somatic cells (middle) and cancer cells (right). Red indicates relative higher density, and blue indicates relative lower density. The yellow dashed lines and percentages highlight the reduction of high linkage disequilibrium (r2 > 0.9) with n=500,000 sampling. (d) Co-localization of MHBs (n=147,888) with known genomic features. (e) Enrichment of MHBs (n=147,888) in known genomic features.

We started with 51 sets of published WGBS data from human primary tissues[49,87], the H1 human embryonic stem cells, *in-vitro*-derived progenitors[13] and human cancer cell lines[88,89]. We also included an in-house-generated WGBS data set from ten adult tissues of one human donor. Across these 61 samples (>2,000× combined genome coverage) we identified a total of ~0.711 billion methylation haplotype

informative reads that covered 58.2% of autosomal CpGs. The uncovered CpG sites were either in regions with low mappability or in CpG-sparse regions in which there were too few CpG sites within the Illumina read pairs to derive informative haplotypes. We partitioned the human genome into blocks of tightly coupled CpG methylation sites (which we refer to as MHBs; **Figure 4-1b**), using a r^2 cutoff of 0.5. We identified 147,888 MHBs at an average size of 95 bp and a minimum of three CpGs per block, which represents ~0.5% of the human genome that tends to be tightly co-regulated on the epigenetic status at the level of single DNA molecules (**Figure 4-2a**,**b**). The majority of CpG sites within the same MHBs were nearly perfectly coupled (r^2 ~ 1.0) regardless of the sample type. We found that the fraction of tightly coupled CpG pairs (r^2 > 0.9; **Figure 4-1c**) slightly decreased over CpG spacing from stem and progenitor cells (94.8%; mostly cultured cells) to somatic cells (91.2%; mixture of primary adult tissues) to cancer cells (87.8%; mixture of CRC tissues and LC cell lines).

Although the WGBS data came from different laboratories, which might have technical differences from batch to batch, we found that that methylation LD extended further over CpG distance in stem and progenitor cells, which is consistent with our previous observations on 2,020 CpG islands[59] for culture cell lines and with another report[90]. Notably, in cancer samples, we observed a reduction of perfectly coupled CpG pairs, which could be related to the pattern of discordant methylation that was recently reported in variable-methylation regions (VMRs)[57,78]. The cancer-specific decayed MHBs were enriched for cancer-related pathways and functions (**Table 4-1**). Nonetheless, the majority of MHBs in cancers still contains tightly coupled CpGs (87.8%), allowing us to harness the pattern for detecting tumors in plasma.

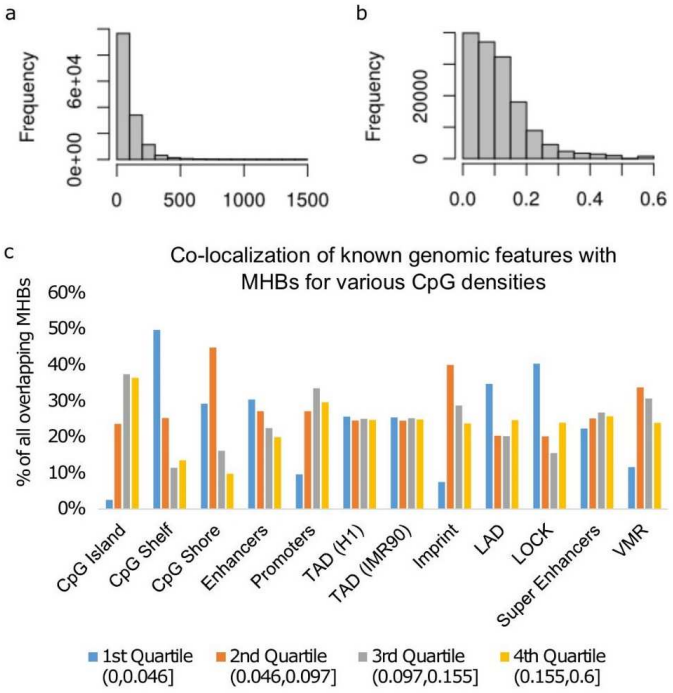


Figure 4-2. Characteristics of MHBs in the human genome

(a) Distribution of MHB sizes. (b) Distribution of MHBs CpG densities (CpGs/bp). (c) Co-localization of known genomic features broken down by CpG density. Note that closed brackets are inclusive.

Table 4-1. Gene ontology analysis to cancer loss linkage regions by GREAT

# Term Name	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed	Hyper Total Genes	Hyper Gene Set Coverage
positive regulation of mRNA catabolic process	0.047	1.63	13	14	0.0013
negative regulation of TGF beta receptor signaling pathway	0.000	1.43	54	66	0.0052
positive regulation of fibroblast proliferation	0.000	1.55	38	43	0.0037
apoptotic mitochondrial changes	0.004	1.43	40	49	0.0039
lens fiber cell differentiation	0.010	1.59	20	22	0.0019
positive regulation of mitochondrion organization	0.016	1.36	42	54	0.0041
regulation of insulin receptor signaling pathway	0.001	1.59	29	32	0.0028
regulation of cellular response to insulin stimulus	0.002	1.50	35	41	0.0034
filopodium assembly	0.039	1.51	19	22	0.0018
positive regulation of macrophage differentiation	0.040	1.75	10	10	0.0010
mature B cell differentiation	0.040	1.75	10	10	0.0010
regulation of release of cytochrome c from mitochondria	0.007	1.50	29	34	0.0028
regulation of monocyte differentiation	0.040	1.75	10	10	0.0010
cellular metabolic compound salvage	0.006	1.53	27	31	0.0026
positive regulation of protein deacetylation	0.016	1.75	12	12	0.0012
stress fiber assembly	0.040	1.75	10	10	0.0010
pyrimidine-containing compound salvage	0.040	1.75	10	10	0.0010
focal adhesion	0.006	1.27	95	131	0.0092
cell-substrate adherens junction	0.007	1.26	97	135	0.0094
Genes related to Wnt-mediated signal transduction	0.013	1.28	65	89	0.0063
Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha)	0.036	1.30	43	58	0.0042
Validated targets of C-MYC transcriptional repression	0.007	1.36	49	63	0.0048
Ceramide signaling pathway	0.044	1.32	36	48	0.0035
Erk1/Erk2 Mapk Signaling pathway	0.033	1.44	23	28	0.0022
Genes involved in RORA Activates Circadian Expression	0.040	1.46	20	24	0.0019
Genes involved in TGF-beta receptor signaling activates SMADs	0.040	1.46	20	24	0.0019
Genes involved in Downregulation of TGF-beta receptor signaling	0.036	1.50	18	21	0.0017
Hypoxia and p53 in the Cardiovascular system	0.006	1.60	21	23	0.0020
Involved in Sema4D induced cell migration and growth-cone collapse	0.040	1.46	20	24	0.0019
IL-2 Receptor Beta Chain in T cell Activation	0.013	1.43	31	38	0.0030
Genes involved in Regulation of IFNG signaling	0.044	1.62	12	13	0.0012
Cell Cycle: G1/S Check Point	0.012	1.50	24	28	0.0023
Regulation of cytoplasmic and nuclear SMAD2/3 signaling	0.041	1.55	15	17	0.0015
Regulation of transcriptional activity by PML	0.041	1.55	15	17	0.0015
Genes involved in Sema4D in semaphorin signaling	0.025	1.45	24	29	0.0023
IL3-mediated signaling events	0.017	1.49	23	27	0.0022

While WGBS data allowed us to unbiasedly identify MHBs across the entire genome, the 61 sets of data did not represent the full diversity of human cell/tissue types. To validate the presence of MHBs in a wider range of human tissues and cultured cells, we examined 101 published reduced representation bisulfite sequencing (RRBS) datasets from the ENCODE project that included cell line and normal tissue samples, as well as 637 published Infinium HumanMethylation450K BeadChip (HM450K) datasets

from the TCGA project that included 11 human cancer tissues. The RRBS datasets were generated with short (36bp) Illumina sequencing reads, greatly limiting the length of methylation haplotypes that can be called. Similarly, Illumina methylation arrays only report average CpG methylation of all DNA molecules in a sample, preventing a methylation linkage disequilibrium analysis. Therefore, we calculated the Pearson's correlation coefficient from methylation levels of adjacent CpGs across different sample sets for block partitioning. Note that the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs (**Figure 4-3a**). Nonetheless, the absence of correlated methylation blocks in these data would invalidate the pattern of MHBs. We identified 23,517 and 2,212 correlated methylation blocks from RRBS and HM450K data respectively, among which 8,920 and 1,258 have significant overlaps with WGBS-defined MHBs. Additionally, we observed significantly higher correlation (r^2) among the CpGs within the MHB regions compared CpG loci outside MHBs in HM450K and RRBS dataset (**Figure 4-3b**), further supporting the block-like organization of local CpG co-methylation across a wide variety of cells and tissues. Taken together, the MHBs that we have identified represent a distinct class of genomic feature where local CpG methylation is established or removed in a highly coordinated manner at the level of single DNA molecules, presumably due to the processive activities of the related enzymes coupled with the local density of CpG dinucleotides.

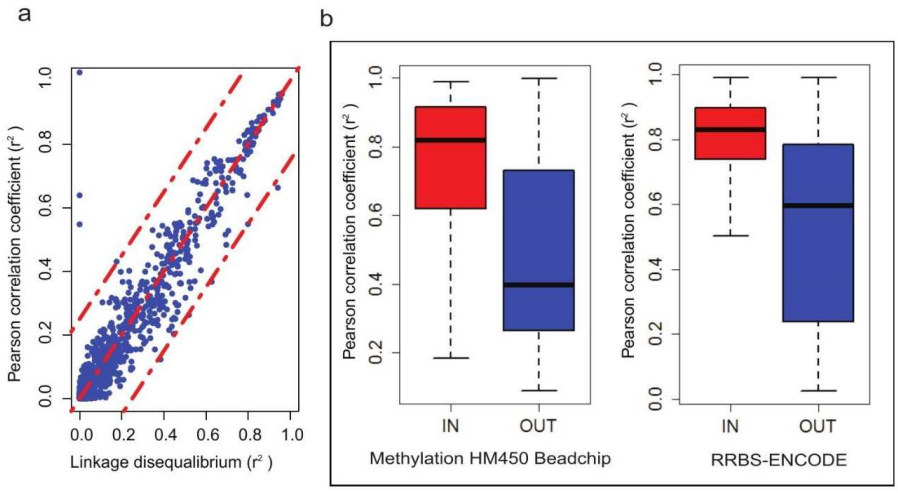


Figure 4-3. Validation of MHBs with TCGA HM450K beadchips and ENCODE RRBS data
(a) Pearson correlatio r^2 versus LD r^2. (b) The Pearson correlation for CpGs in RRBS and HM450K data were significantly higher in regions overlapping with MHBs compared with the CpGs without overlapping with MHBs. IN denotes RRBS or HM450K CpGs within MHBs. OUT denotes RRBS or HM450K regions beyond MHBs.

Co-localization of MHBs with known regulatory elements

The MHBs established by the WGBS data represent a distinct type of genomic feature that partially overlaps with multiple known genomic elements (**Figure 4-1d**). Among all of the MHBs, 60,828 (41.1%) were located in intergenic regions, whereas 87,060 (58.9%) regions were located in transcribed regions. These MHBs were significantly enriched (*p-value* < 1.0×10^{-6}) in enhancers, super-enhancers, promoters, CpG islands and imprinted genes. In addition, we observed a modest depletion in the lamina-associated domains (LADs)[91] and the large organized chromatin Lys9 modifications (LOCK) regions[92], as well as a modest enrichment in defined topologically associated domains (TADs)[93]. Notably, we observed a strong (26-fold) enrichment in VMRs (**Figure 4-1e**), suggesting that increased epigenetic variability in a cell population or tissue can be coordinated locally among hundreds of thousands of genomic regions[94]. We further examined a subset of MHBs that did not overlap with CpG islands and observed a consistent enrichment pattern (**Figure 4-1e**, **4-2c**), suggesting

that local CpG density alone does not account for the enrichment. Previous studies on mice and humans[63,74] demonstrated that dynamically methylated regions are associated with regulatory regions, such as enhancer-like regions marked by acetylation on Lys27 of histone H3 (H3K27ac) and transcription-factor-binding sites. In publicly available histone-mapping data for human adult tissues, we found co-localization of MHBs with marks for active promoters (trimethylated Lys4 on histone H3 (H3K4me3) with H3K27ac) but not for active enhancers[95] (no peak for H3K4me1) (**Figure 4-4**). We found that enhancers tended to overlap with CpG-sparse MHBs, whereas the co-localization with super-enhancers was independent of CpG density (**Figure 4-2c**). Therefore, MHBs probably capture the local coherent epigenetic signatures that are directly or indirectly coupled to transcriptional regulation.

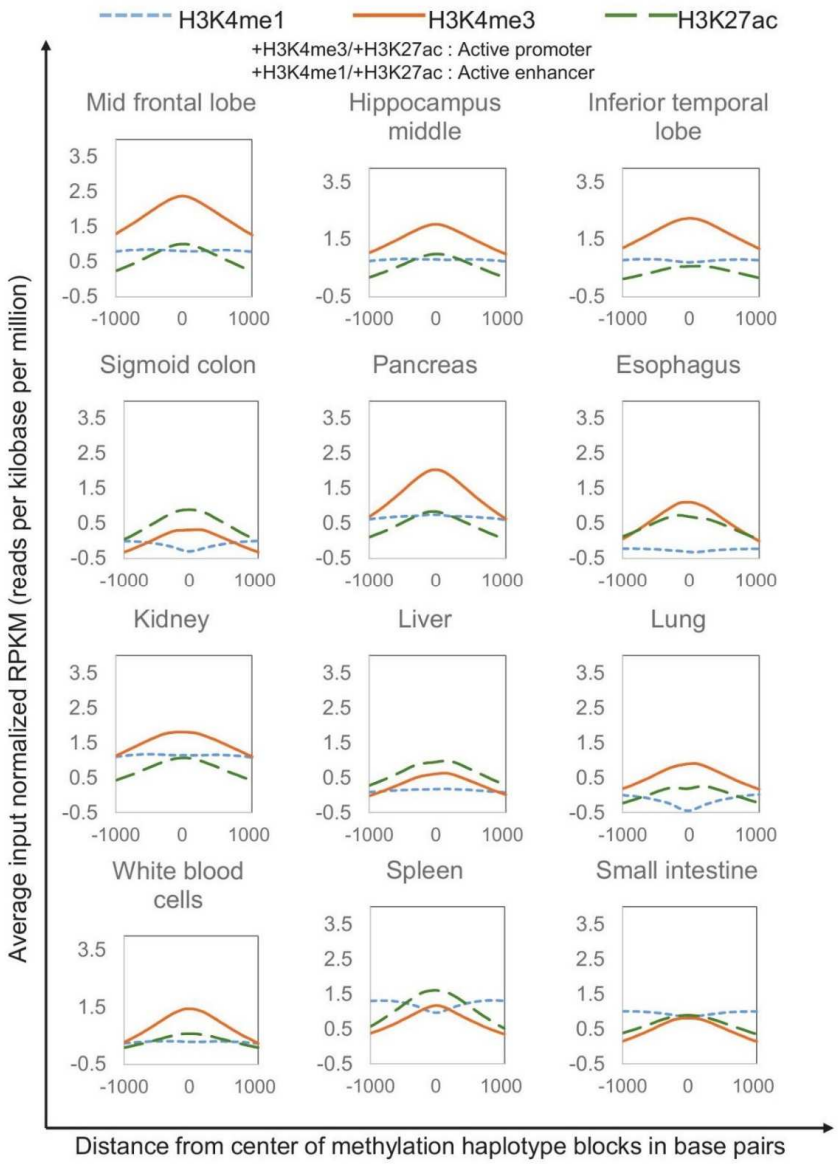
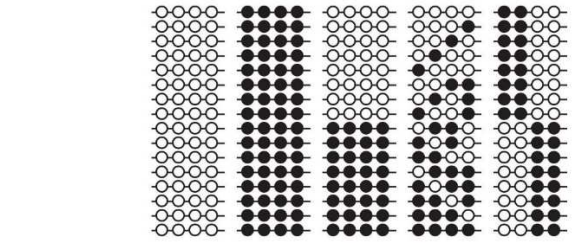


Figure 4-4. Profiles of H3K27ac, H3K4me3 and H3K4me1 over methylation haplotype blocks for 12 human adult tissue types

X-axis denote the distances from the centers of MHBs (+/- 1000 bp) and y-axis denotes the average reads density in RPKM (input normalized reads per kilobase per million). Epigenomics Roadmap histones data were downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/).

Block-level analysis using methylation haplotype load

To enable quantitative analysis of the methylation patterns within individual MHBs across many samples, we needed a single metric to define the methylated pattern of multiple CpG sites within each block. Ideally this metric should not only be a function of the average methylation level for all of the CpG sites in the block, but it should also be able to capture the pattern of co-methylation on single DNA molecules. Therefore, we defined MHL as the weighted mean of the fraction of fully methylated haplotypes and substrings at different lengths (i.e., all possible substrings). As compared to the other metrics used in the literature (methylation level, methylation entropy, epi-polymorphism and haplotype counts), the MHL is capable of distinguishing blocks that have the same average levels of methylation but various degrees of coordinated methylation (**Figure 4-5**). In addition, the MHL is bounded between 0 and 1, which allows for direct comparison of different regions across many data sets.



Methylation Frequency	0%	100%	50%	50%	50%
Methylation Entropy	0	0	0.25	1.00	0.25
Epi-polymorphism	0	0	0.50	0.375	0.50
Haplotypes	1	1	2	16	2
MHL($w_i = i$)	0	1	0.50	0.1625	0.12

Figure 4-5. Comparison of methylation haplotype load with four other metrics used in the literature.
Five patterns of methylation haplotype combinations (schematic) are used to illustrate the differences between methylation frequency, methylation entropy, epi-polymorphism and MHL. MHL is the only metric that can discriminate all five patterns.

We next asked whether treating MHBs as individual genomic features and performing quantitative analysis based on the MHL would provide an advantage over

previous approaches that used individual CpG sites or weighted (or unweighted) averaging of multiple CpG sites in certain genomic windows. Therefore, we clustered 65 WGBS data sets (including four additional colon and lung cancer WGBS sets[89]) from human solid tissues on the basis of the MHL. Principle component analysis (PCA) on all MHBs yielded a clustering of samples with same tissue of origin (**Figure 4-6**). Unsupervised clustering with the 15% most-variable MHBs showed that, regardless of the data sources, samples of the same tissue origin clustered together (**Figure 4-7a**), whereas cancer samples and stem cell samples showed patterns distinct from those of human adult tissues. To identify a subset of MHBs for effective clustering of human somatic tissues, we constructed a tissue specific index (TSI) for each MHB. Random Forest based feature selection identified a set of 1,360 tissue-specific MHBs that can predict tissue type at an accuracy of 0.89 (95%CI: 0.84-0.93), despite the fact that several tissue types share rather similar cell compositions (i.e. muscle vs. heart). Using this set of MHBs, we compared the performance between MHL, average methylation fraction in the MHL regions (AMF) and all individual CpG methylation fraction (IMF). MHL and the average methylation provided similar tissue specificity, while MHL has a lower noise (background noise: 0.29, 95%CI: 0.23-0.35) compared with average methylation (background noise: 0.4, 95%CI: 0.32-0.48). Clustering based on individual CpGs in the blocks has the worst performance that might be due to higher biological or technical viability of individual CpG sites (**Figure 4-7c**). Thus block-level analysis based on MHL is advantageous over single CpG or local averaging of multiple CpG sites in distinguishing tissue types from regions of coupled CpG methylation and heterogeneity.

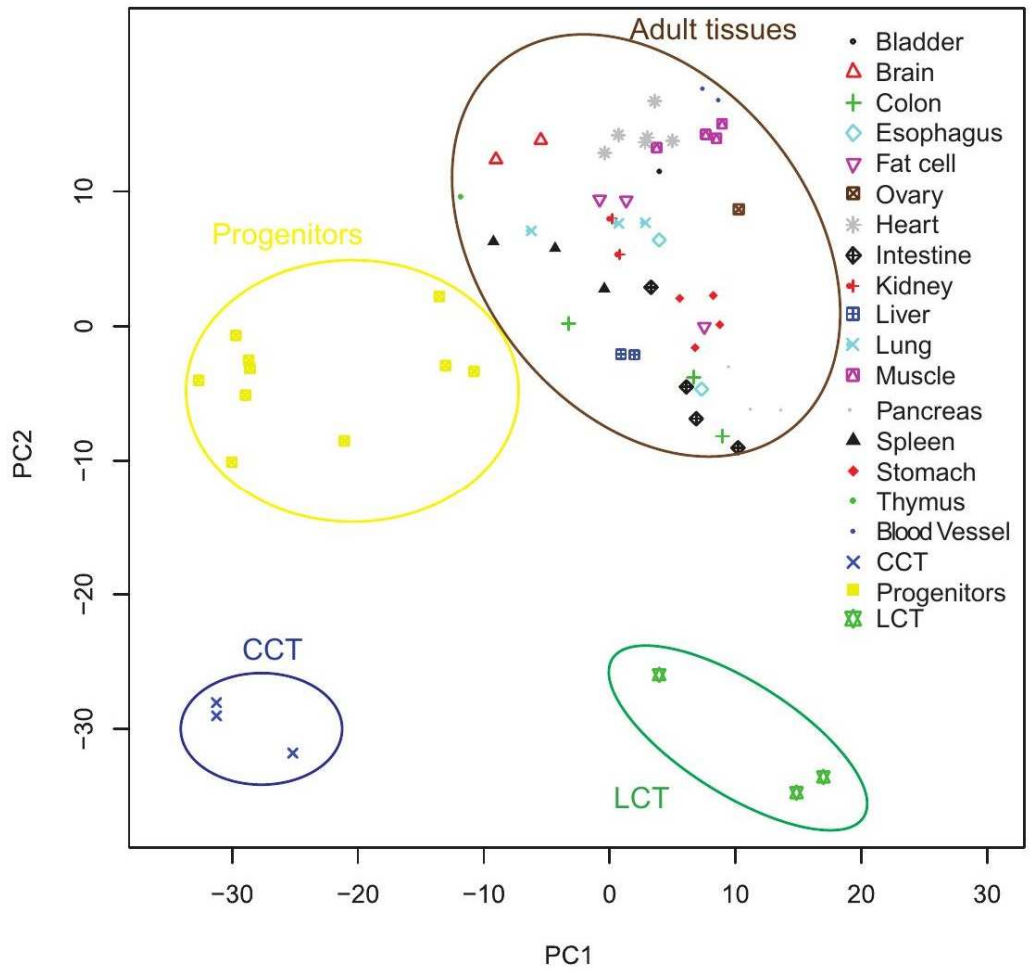


Figure 4-6. PCA of human tissues and cells based on methylation haplotype loads in MHB regions.

Ten adult tissues WGBS were from this study and others were from 5 published studies.

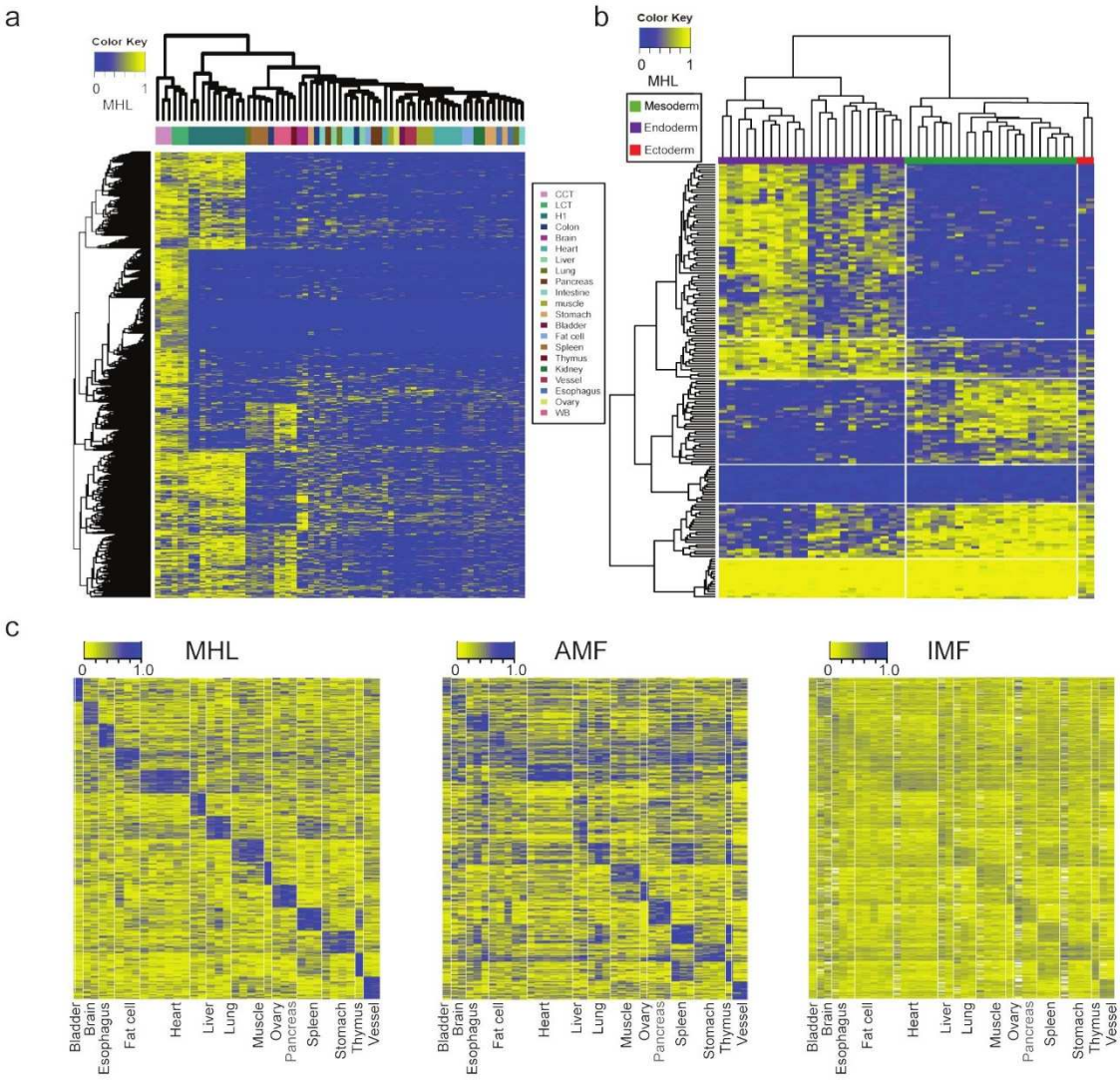


Figure 4-7. Tissue clustering based on methylation haplotype load

(a) MHL-based unsupervised clustering of human tissues using the top 15% most-variable regions. Color bar indicates the MHL value. (b) Supervised clustering of germ-layer-specific MHBs. (c) Comparison of cluster performance to different samples using matric MHL, AMF and IMF. MHL exhibits better signal-to-noise ratio than AMF and IMF for sample clustering.

The human adult tissues that we used have various degrees of similarity among each other. We hypothesized that this is primarily defined by their developmental lineage and that the related MHBs might reveal epigenetic insights relevant to germ-layer specification. We searched for MHBs that had differential MHLs among the data sets from the three germ layers. In total we identified 114 ectoderm-specific MHBs (99 hypermethylated and 15 hypomethylated), 75 endoderm-specific MHBs (58 hypermethylated and 17 hypomethylated) and 31 mesoderm-specific MHBs (9 hypermethylated and 22 hypomethylated). Cluster analysis based on layer-specific MHBs showed the expected clustering among tissues of the same lineage (**Figure 4-7b**). We speculated that some of these MHBs might capture binding events of transcription factors (TFs) specific to the developmental germ layers. We observed patterns of TF binding to layer-specific MHBs that overlapped with ENCODE reported TF-binding events[83] (**Figure 4-8**). For layer-specific MHBs with low MHLs, we identified 53 TFs in mesoderm-specific MHBs, 71 TFs in endoderm-specific MHBs and 2 TFs in ectoderm-specific MHBs. Gene ontology analysis showed that mesoderm-specific TFs binding events have negative-regulator activity, whereas endoderm-specific TFs binding events have positive-regulator activity. For layer-specific MHBs with a high MHL, we identified 38 TFs in mesoderm-specific MHBs, 102 TFs in endoderm-specific MHB and 145 TFs in ectoderm-specific MHBs. Notably, tissues in the ectoderm and endoderm lineage shared few bounded TFs, whereas mesoderm tissue shared multiple groups of TFs with the ectoderm and endoderm tissues. We identified two endoderm-specific high-MHL regions, which are associated with the transcription factors ESRRA (also known as ERR1) and NANOG. This is consistent with a previous finding that mouse embryonic stem cells differentiated spontaneously into visceral and parietal endoderm after knocking out *Nanog*[96]. The low-MHL regions shared by the mesoderm and endoderm

might have regulatory functions in the fate commitment toward multiple tissues, whereas the ectoderm-specific high-MHL regions might induce ectoderm development by suppressing the path toward cells of the immune lineage (**Figure 4-8**). These observations are indicative of two distinctive 'push' and 'pull' mechanisms in the transition of cell states that have been harnessed for the induction of pluripotency by overexpressing lineage specifiers[97].

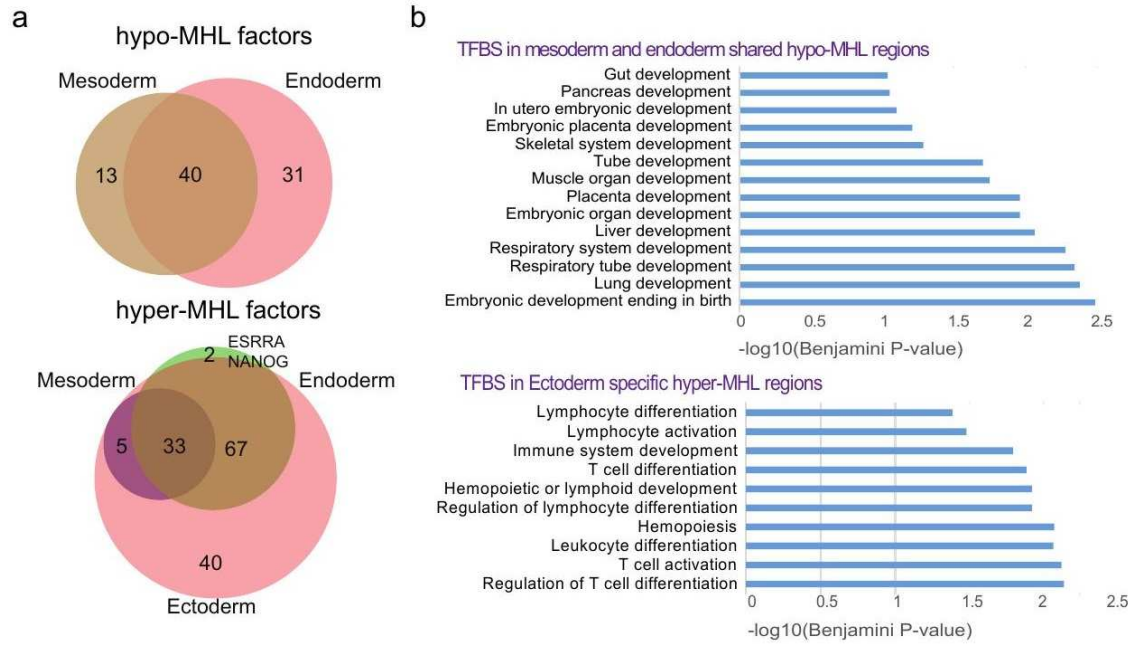


Figure 4-8. Distinct patterns of functional enrichment for TFBS associated with layer-specific MHBs.

(a) Venn diagrams of transcription factors (TF) with binding sites associated with layer specific hypo- or hyper- MHL regions. (b) Functional enrichment analysis of associated TFBS using GREAT (http://bejerano.stanford.edu/great/public/html/).

Methylation-haplotype-based analysis of circulating cfDNA

A unique aspect of methylation-haplotype analysis is that the pattern of co-methylation, especially within MHBs, is robust in capturing low-frequency alleles among a heterogeneous population of molecules or cells, in the presence of biological noise or technical variability, such as incomplete bisulfite conversion or sequencing errors. To explore potential clinical applications, we next focused on the methylation-haplotype

analysis of cfDNA from healthy donors and patients with cancer, in which low fractions of DNA molecules released from tumor cells and potentially carry epigenetic signatures distinct from those of white blood cells. We isolated cfDNA from the plasma of 75 healthy individuals (NCP), 29 patients with lung cancer (LCP) and 30 patients with colorectal cancer (CCP). Owing to the limited amounts of available DNA, we performed single-cell RRBS (sc-RRBS)[98] and obtained an average of 13 million paired-end 150-bp reads per sample. On average, 57.7% of WGBS-defined MHBs were covered in our RRBS data set from the clinical samples.

We sought to detect the presence of tumor specific signatures in the plasma samples, using methylation haplotypes identified from tumor tissues as the reference and normal samples as the negative controls. For five lung cancer plasma samples and five colorectal cancer plasma samples, we also obtained matched primary tumor tissues, and generated RRBS data (30 million reads per sample) from 100ng of tumor genomic DNA. We focused on MHBs with low MHL (i.e. genomic regions that have low or no methylation) in the blood, and asked whether we can detect cancer-associated highly methylated haplotypes (caHMH). We required that such haplotypes were present only in the tumor tissues and the matched plasma from the same patient, but not in whole blood or any other non-cancer samples. We considered these highly confident tumor signature in circulating DNA. We detected caHMH in all cancer patient plasma samples ("caHMH-2", Average=36, IQR=17, **Table 4-2**). These HMHs were associated with 183 genes, some of which are known to be aberrantly methylated in human cancers such as *WDR37*, *VAX1*, *SMPD1*. Next, we extended the analysis to 49 additional cancer plasma samples that have no matched tumor samples, using 65 normal plasmas as the background. On average, 60 (IQR=31) caHMH were identified for each cancer plasma sample (**Table 4-3**). Interestingly, a significant fraction (35%) of caHMH called on

matched tumor-plasma pairs were also detected the expanded set of cancer patient plasma samples.

Table 4-2. Cancer associated High Methylation Haplotype based on matched plasma-tumor tissue samples

Patient	# caHMH candidates	# caHMH-1	**# caHMH-2***	# caHMH-3
CRC-P-1	1885	526	**37**	14
CRC-P-2	1257	340	**20**	9
CRC-P-3	3630	880	**35**	19
CRC-P-4	1700	509	**55**	21
CRC-P-5	2062	614	**21**	10
LC-P-1	2065	550	**16**	5
LC-P-2	2320	571	**36**	15
LC-P-3	1959	566	**15**	11
LC-P-4	2068	658	**97**	46
LC-P-5	1799	524	**30**	15

* caHMH-2 were used in cancer DNA fragment estimation in Figure 4-5. caHMH-1: Cancer associated HMH defined as shared by primary tumor tissue and paired plasma while not in whole blood (WB).caHMH-2: Cancer associated HMH defined as shared by primary tumor tissue and paired plasma while not in normal plasma and WB.caHMH-3: Cancer associated HMH defined as shared by primary tumor tissue and paired plasma while not in normal plasma, WB, and normal tissues.

Table 4-3. Cancer associated HMH in all plasma samples

	# caHMH identified
CRC-P-10	48
CRC-P-11	31
CRC-P-12	247
CRC-P-13	38
CRC-P-14	27
CRC-P-15	89
CRC-P-16	34
CRC-P-17	303
CRC-P-18	21
CRC-P-19	21
CRC-P-20	32
CRC-P-21	327
CRC-P-22	85
CRC-P-23	56
CRC-P-24	70
CRC-P-25	25
CRC-P-26	29
CRC-P-27	131
CRC-P-28	26
CRC-P-29	47
CRC-P-30	21
CRC-P-6	21
CRC-P-7	55
CRC-P-8	28
CRC-P-9	43
LC-P-10	37
LC-P-11	38
LC-P-12	23
LC-P-13	18
LC-P-14	29
LC-P-15	24
LC-P-16	24
LC-P-17	34
LC-P-18	243
LC-P-19	40
LC-P-20	43
LC-P-21	33
LC-P-22	62
LC-P-23	36
LC-P-25	15
LC-P-26	10
LC-P-27	96
LC-P-28	18
LC-P-29	15
LC-P-30	51
LC-P-6	42
LC-P-7	40
LC-P-8	63
LC-P-9	46

Next we quantified the tumor load in cancer plasma samples, using non-negative decomposition with quadratic programming, on the RRBS data from primary cancer biopsies (LC & CRC) and WGBS data from 10 normal tissues. We estimated that a predominant fraction, 72.0% (95% CI:0.659-0.782) in the cancer and normal plasma were contributed by white blood cells, which is consistent with the levels reported recently based on shallow whole genome bisulfite sequencing (69.4%)[31]. Primary tumor and normal tissue-of-origin contributed at the similar level of 2.3% (95% CI: 0.4%-4.2%) and 3.0% (95% CI:1.2%-4.4%). In contrast, we applied the similar analysis to normal plasma, and found only residual tumor contributions (0.17% for CRC and 1.0% LC) to normal plasma, which were significantly lower (*p-value* =3.4×10^{-5} & 5.2×10^{-10} for CRC and LC, respectively) than cancer plasma. We also found that 76.7% plasma samples from CRC patients and 89.6% from LC patients had detectible contribution from tumor tissues while only 13% and 26% normal plasmas have certain (low) tumor contribution (**Figure 4-9**). Therefore, circulating cell-free DNA contains a relatively stable fraction of molecules released from various normal tissues, whereas in cancer patients tumor cells released DNA molecules that can be more abundant than normal tissues (**Table 4-4**).

Table 4-4. Deconvolution of plasma samples to 10 normal tissues, lung cancer tissues(LCT), and colon cancer tissues(CCT)

	Brain	CCT	Colon	Esophagus	Heart	Intestine	Kidney	LCT	Liver	Lung	Stomach	WB
CCP	0.027	0.015	0.019	0.036	0.030	0.035	0.031	0.030	0.145	0.046	0.044	0.543
NP	0.015	0.002	0.002	0.001	0.037	0.010	0.013	0.010	0.056	0.044	0.003	0.808
LCP	0.045	0.013	0.057	0.046	0.047	0.041	0.048	0.035	0.095	0.044	0.042	0.488

Average values from only plasma samples with WB > 0.3.

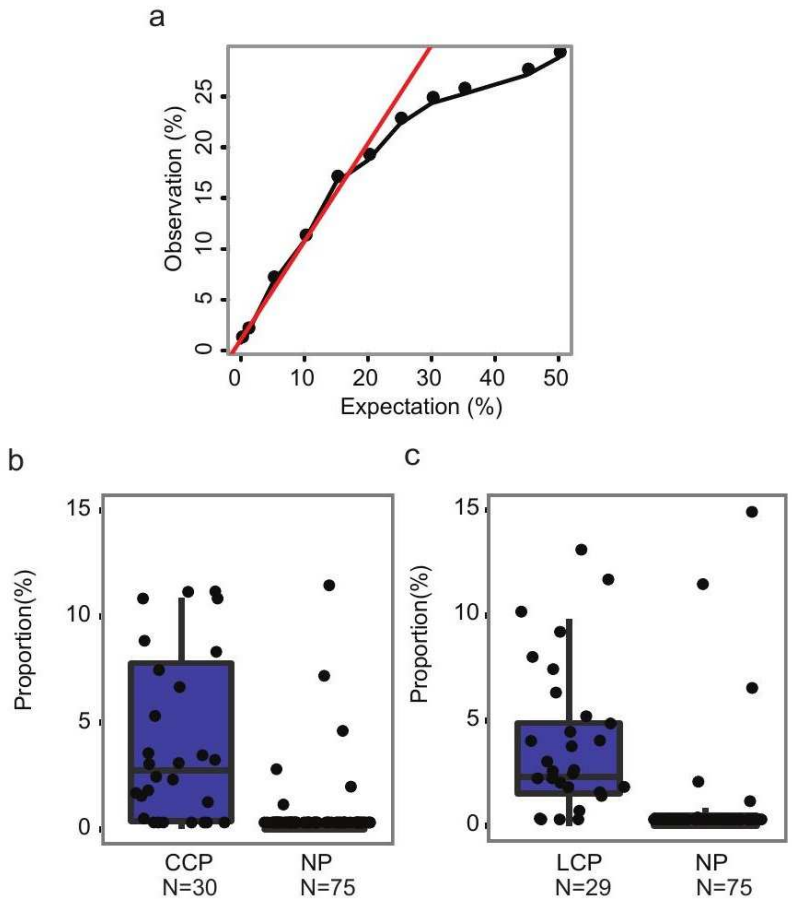


Figure 4-9. Deconvolution of cancer and normal plasma samples using non-negative decomposition with quadratic programming.

(a) Deconvolution accuracy as a function of tumor fraction using simulated data. (b) Cancer DNA proportions estimated by deconvolution of plasma samples using CCT or LCT as the tumor reference.

We next asked whether we can identify a small subset of MHBs among all the RRBS targets that have significantly higher levels of MHL in cancer plasma than in normal plasma. We found 81 and 94 MHBs with significantly higher MHL for colorectal and lung cancer. Some of these regions (such as *HOXA3*) have been reported to be aberrantly methylated in lung cancer and colorectal cancer. Using these MHBs as markers, the diagnostic sensitivity is 96.7% and 93.1% for colorectal cancer and lung cancer at the specificity 94.6% and 90.6%. As a comparison, we also performed a

prediction based on average 5mC methylation level within these MHB regions, or based on genome-wide single CpG sites. MHL was found to be superior to average 5mC methylation level (sensitivity of 90.0% and 86.2%; specificity of 89.3% and 90.6% for CRC and lung cancer) and methylation signal of individual CpG site (sensitivity of 89.6% and 80.6%; specificity of 89.3% and 92.0%).

We then sought to use the information from normal human tissues, primary tumor biopsies and cancer cell lines to improve the detection of ctDNA. We started by selecting a subset of MHBs that show high MHL (>0.5) in primary cancer biopsies and low MHL (<0.1) in whole blood, then clustered these MHBs into three groups based on the MHL in all normal and cancer plasma, as well as cancer and normal tissues (**Figure 4-10**). We identified a subset (Group II) of MHBs that have high MHL in cancer tissues and low MHLs in normal tissues. Cancer plasma showed significantly higher MHL in these regions than normal plasma (P=1.4×10^{-12} and 6.2×10^{-8} for CRC and LC, respectively). By computationally mixing the sequencing reads from cancer tissues and whole blood samples (WBC), we created synthetic admixtures at various levels of tumor fraction. We found that MHL is 2-5 folder higher than the methylation level of individual CpG sites across the full range of tumor fractions. Remarkably, MHL provides additional gain of signal-to-noise ratio (mean divided by standard deviation) compared with AMF as the fraction of tumor DNA decreases below 10%, which is typical for clinical samples (**Figure 4-10c**). We then took the individual plasma data sets, and predicted the tumor fraction based on the MHL distribution established by computational mixing (**Figure 4-10a-b)**. Except for a small number (N<5) of outliers, we observed significantly higher average MHL in cancer plasma than in normal plasma (**Table 4-5**, **Table 4-6**). Note that all Group II MHBs were selected without using any information from the plasma samples, and hence they should be generally applicable to other plasma samples.

Interestingly, we also found that the estimated tumor DNA fraction were positive correlated with normalized cfDNA yield from the cancer patients (P<0.000023, **Figure 4-11**).

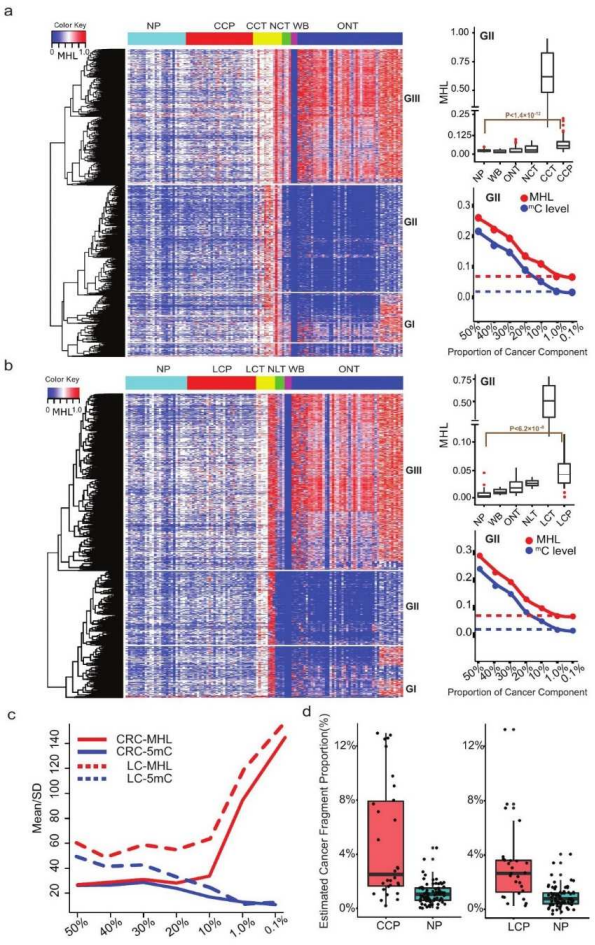


Figure 4-10. Quantitative estimation of the proportion of DNA derived from cancer cells in cell-free DNA, using the MHL of informative MHBs

(a,b) Left, heat maps showing the different patterns of MHL in patients with colorectal cancer (a) or lung cancer (b), as compared to that in healthy individuals (NP). GII regions have high MHL values in tissues (MHL > 0.5) and plasma from patients with cancer and low MHL values in WB and healthy tissues (MHL < 0.1). Bar plots show average MHL values in different groups of samples. MHLs in the plasma of patients with colorectal cancer (CCP) or lung cancer (LCP) and in the plasma of healthy individuals (NP) were compared with a two-tailed Student's t-test. NCT denotes healthy colon tissues, NLT denotes healthy lung tissues, and ONT denotes other healthy tissues. (c) Comparison between signal-to-noise ratio of MHL and 5mC changes as the deduction of tumor DNA fraction. MHL has higher signal-to-noise ratio (Mean/SD ratio) than individual 5mC levels as tumor fraction decreases. x axis shows the tumor fraction in synthetic mixtures. 30 CRC and 29 LC samples were involved in the analysis (d) Estimation of the cancer DNA proportions in plasma samples (30 CCP, 29 LC and 75 NP).

Table 4-5. Relationship between Group II average MHL and cfDNA yield for cancer patient

ID	Type	MHL	Prediction	Plasma volume (uL)	Amount used for RRBS input (ng)	Normalized yield per 1 mL (ng)
CRC.P.001	CCP	0.0232	Colon	430	1.0	13.95
CRC.P.002	CCP	0.0246	Spleen	540	1.0	61.11
CRC.P.003	CCP	0.0080	Colon	220	1.0	214.09
CRC.P.004	CCP	0.0043	Liver	550	1.0	48.49
CRC.P.005	CCP	0.0127	Lung	665	1.0	21.97
CRC.P.006	CCP	0.0069	Colon	650	1.0	11.49
CRC.P.007	CCP	0.0273	Colon	700	1.0	10.59
CRC.P.008	CCP	0.0264	Spleen	690	1.0	9.83
CRC.P.009	CCP	0.0009	Colon	520	1.0	23.54
CRC.P.010	CCP	0.0170	Colon	700	1.0	14.27
CRC.P.011	CCP	0.0203	Colon	231	1.5	31.95
CRC.P.012	CCP	0.1361	Colon	670	1.5	13.97
CRC.P.013	CCP	0.0579	Colon	690	1.5	45.30
CRC.P.014	CCP	0.0265	Colon	585	1.5	24.41
CRC.P.015	CCP	0.0690	Colon	610	1.5	9.54
CRC.P.016	CCP	0.0238	Colon	650	1.5	18.09
CRC.P.017	CCP	0.1349	Colon	480	1.5	16.63
CRC.P.018	CCP	0.0251	Colon	425	1.5	22.31
CRC.P.019	CCP	0.0231	Colon	650	1.5	18.18
CRC.P.020	CCP	0.0183	Colon	641	1.5	8.42
CRC.P.021	CCP	0.1340	Colon	670	1.5	17.01
CRC.P.022	CCP	0.1335	Liver	460	1.5	84.00
CRC.P.023	CCP	0.0248	Colon	900	1.5	29.27
CRC.P.024	CCP	0.0275	Colon	725	1.5	15.14
CRC.P.025	CCP	0.0139	Colon	150	1.5	46.40
CRC.P.026	CCP	0.0376	Colon	550	1.5	17.89
CRC.P.027	CCP	0.0715	Colon	55	1.5	92.73
CRC.P.028	CCP	0.0153	Colon	940	1.5	130.21
CRC.P.029	CCP	0.0133	Colon	940	1.5	10.02
CRC.P.030	CCP	0.0246	Colon	405	1.5	31.70
LC.P.001	LCP	0.0057	Lung	475	1.0	43.71
LC.P.002	LCP	0.0230	Lung	350	1.0	97.71
LC.P.003	LCP	0.0350	WBC	398	1.0	15.68
LC.P.004	LCP	0.0639	Lung	325	1.0	18.65
LC.P.005	LCP	0.0901	Lung	440	1.0	75.68
LC.P.006	LCP	0.0025	Lung	345	1.0	20.26
LC.P.007	LCP	0.0038	Lung	320	1.0	21.47
LC.P.008	LCP	0.0119	Lung	303	1.0	16.53
LC.P.009	LCP	0.0366	Lung	330	1.0	21.00
LC.P.010	LCP	0.0129	Lung	500	1.0	12.18
LC.P.011	LCP	0.0719	Lung	595	1.5	21.28
LC.P.012	LCP	0.0178	Lung	535	1.5	8.75
LC.P.013	LCP	0.0108	Lung	630	1.5	15.33
LC.P.014	LCP	0.0152	Lung	600	1.5	18.50
LC.P.015	LCP	0.0382	Liver	455	1.5	21.23
LC.P.016	LCP	0.0022	Lung	630	1.5	14.00
LC.P.017	LCP	0.0121	WBC	355	1.5	91.77
LC.P.018	LCP	0.5147	Lung	430	1.5	65.67
LC.P.019	LCP	0.0342	Lung	760	1.5	17.37
LC.P.020	LCP	0.0160	Lung	550	1.5	40.69
LC.P.021	LCP	0.0834	Lung	620	1.5	17.81
LC.P.022	LCP	0.0379	Lung	385	1.5	59.84
LC.P.023	LCP	0.0003	Lung	540	1.5	88.89
LC.P.025	LCP	0.0336	Lung	700	1.5	12.51
LC.P.026	LCP	0.0015	Lung	690	1.5	11.48
LC.P.027	LCP	0.1629	Lung	700	1.5	23.57
LC.P.028	LCP	0.0103	Lung	780	1.5	40.77
LC.P.029	LCP	0.0303	Lung	690	1.5	12.00
LC.P.030	LCP	0.0173	Lung	398	1.5	174.87

Table 4-6. Relationship between Group II average MHL and cfDNA yield in healthy controls

ID	Type	MHL	Prediction	Plasma volume (µL)	Amount used for RRBS input (ng)	Normalized yield per 1 mL (ng)
NC.P.001	NP	0.0001	WBC	1000	1.0	12.31
NC.P.002	NP	0.0010	WBC	1000	1.0	13.65
NC.P.003	NP	0.0001	WBC	1000	1.0	22.60
NC.P.005	NP	0.0313	WBC	1000	1.0	6.55
NC.P.006	NP	0.0000	Liver	1000	1.0	6.22
NC.P.007	NP	0.0002	WBC	1000	1.0	10.29
NC.P.008	NP	0.0000	Lung	1000	1.0	5.92
NC.P.009	NP	0.0078	Lung	1000	1.0	8.55
NC.P.012	NP	0.0010	WBC	1000	1.0	8.15
NC.P.013	NP	0.0313	WBC	1000	1.0	8.40
NC.P.014	NP	0.0000	WBC	1000	1.0	7.85
NC.P.015	NP	0.0002	WBC	1000	1.0	5.90
NC.P.016	NP	0.0002	WBC	1000	1.0	7.00
NC.P.017	NP	0.0001	WBC	1000	1.0	6.30
NC.P.018	NP	0.0000	WBC	1000	1.0	9.45
NC.P.019	NP	0.0039	WBC	1000	1.0	5.25
NC.P.020	NP	0.0078	WBC	1000	1.0	6.40
NC.P.021	NP	0.0010	WBC	1000	1.0	10.25
NC.P.022	NP	0.0010	Spleen	1000	1.0	9.00
NC.P.023	NP	0.0078	WBC	1000	1.0	7.65
NC.P.024	NP	0.0078	WBC	1000	1.0	7.00
NC.P.025	NP	0.0001	WBC	1000	1.0	4.15
NC.P.026	NP	0.0002	Liver	1000	1.0	3.60
NC.P.027	NP	0.0000	Spleen	1000	1.0	6.20
NC.P.029	NP	0.0000	WBC	1000	1.0	5.45
NC.P.030	NP	0.0143	WBC	1000	1.0	4.95
NC.P.031	NP	0.0125	WBC	1500	10.0	14.66
NC.P.032	NP	0.0144	WBC	1400	10.0	11.55
NC.P.033	NP	0.0324	WBC	1500	10.0	13.80
NC.P.034	NP	0.0255	WBC	1500	10.0	175.00
NC.P.035	NP	0.0315	WBC	1250	10.0	12.74
NC.P.036	NP	0.0197	WBC	1500	10.0	15.54
NC.P.037	NP	0.0506	WBC	1400	10.0	13.76
NC.P.038	NP	0.0212	WBC	1100	10.0	35.73
NC.P.039	NP	0.0165	WBC	1350	10.0	20.60
NC.P.040	NP	0.0181	WBC	950	10.0	15.60
NC.P.041	NP	0.0140	WBC	1200	10.0	33.25
NC.P.043	NP	0.0053	WBC	1150	10.0	9.60
NC.P.044	NP	0.0186	WBC	1350	10.0	20.96
NC.P.045	NP	0.0011	WBC	1300	10.0	119.31
NC.P.046	NP	0.0148	WBC	1350	10.0	12.33
NC.P.047	NP	0.0065	WBC	800	10.0	27.71
NC.P.048	NP	0.0444	WBC	1000	10.0	24.84
NC.P.049	NP	0.0004	WBC	1100	10.0	15.38
NC.P.050	NP	0.0080	WBC	850	10.0	18.99
NC.P.051	NP	0.0190	WBC	1350	10.0	15.11
NC.P.052	NP	0.0130	WBC	1100	10.0	16.64
NC.P.053	NP	0.0053	WBC	1450	10.0	23.38
NC.P.054	NP	0.0076	WBC	1050	10.0	28.23
NC.P.055	NP	0.0088	WBC	1450	10.0	19.74
NC.P.056	NP	0.0051	WBC	1300	10.0	47.31
NC.P.057	NP	0.0478	WBC	1300	10.0	74.54
NC.P.058	NP	0.0014	WBC	1300	9.9	327.69
NC.P.059	NP	0.0013	WBC	1350	10.0	12.29
NC.P.060	NP	0.0421	WBC	1150	10.0	24.31
NC.P.061	NP	0.0180	WBC	1400	10.0	31.50
NC.P.062	NP	0.0019	WBC	1300	10.0	48.46
NC.P.063	NP	0.0226	WBC	1350	10.0	50.00
NC.P.064	NP	0.0215	WBC	1350	10.0	41.11
NC.P.065	NP	0.0053	WBC	1200	9.9	8.20
NC.P.066	NP	0.0079	WBC	1300	10.0	70.85
NC.P.067	NP	0.0049	WBC	1350	10.0	25.78
NC.P.068	NP	0.0075	WBC	1350	10.0	9.47
NC.P.069	NP	0.0081	WBC	1350	10.0	8.11
NC.P.070	NP	0.0086	WBC	1350	10.0	13.27
NC.P.071	NP	0.0045	WBC	1350	10.0	16.00
NC.P.072	NP	0.0059	WBC	1200	10.0	17.83
NC.P.073	NP	0.0174	WBC	1450	10.0	24.41
NC.P.074	NP	0.0025	WBC	1250	10.0	24.72
NC.P.075	NP	0.0073	Lung	1300	10.0	70.38
NC.P.076	NP	0.0151	WBC	1450	10.0	17.44
NC.P.077	NP	0.0031	WBC	1400	10.0	10.65
NC.P.078	NP	0.0323	WBC	1300	10.0	10.73
NC.P.079	NP	0.0422	WBC	1250	10.0	11.28
NC.P.080	NP	0.0192	WBC	1400	10.0	14.0

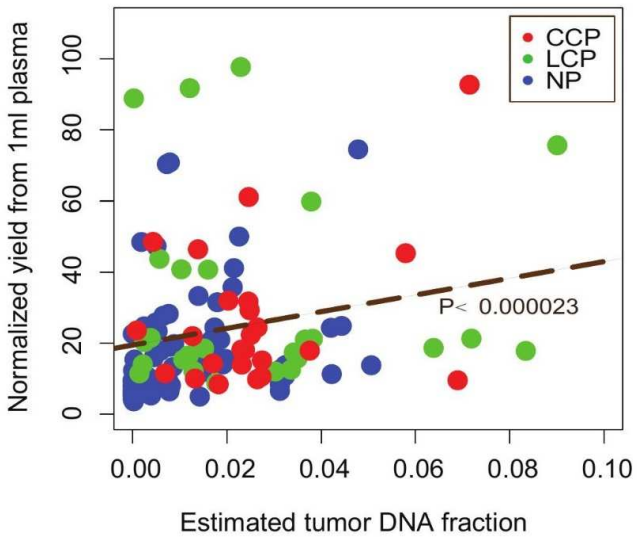


Figure 4-11. Estimated tumor fraction in plasma correlated with the normalized yield of DNA extraction from plasma
Plasma yields were from normal controls, lung cancer patients, and colorectal cancer patients.

Recent studies[30,31,99] have demonstrated that epigenetic information imbedded in cfDNA has the potential for predicting a tumor's tissue of origin. Consistently, we found that tissue-of-origin derived methylation haplotypes were the most abundant fraction in cancer plasma (**Table 4-4**). Here we asked whether a MHL-based framework and a set of targets derived from whole genome data would allow us to predict tissue-of-origin with quantifiable sensitivity and specificity, which is crucial for future clinical applications. We compiled 43 WGBS and RRBS data sets for 10 human normal tissues that have high cancer incident rate, and identified a set of 2,880 tissue-specific MHBs as the candidates. We then used these tissue-specific MHBs or subsets to predict the tissue-of-origin for the cancer plasma sample. Although we found a large number of tissue-of-origin specific MHBs that have low MHL in normal plasma (**Figure 4-12a**), the multiclass prediction based on random forest yielded very limited power, most likely due to the high diversity of the tissue classes (N=10). We then adopted an alternative approach by counting the total number of tissue-specific MHBs in the plasma samples and comparing

with all other tissues, in order to infer the most probable tissue-of-origin. At the cutoff of minimal 10 tissue-specific methylated haplotypes per tissue type, we observed an average 90% accuracy for mapping a data set from the primary tissue to its tissue type (**Figure 4-12b**). We then applied this method to the full set of plasma data from 59 cancer patients and 75 normal individuals, and achieved an average prediction accuracy of 82.8%, 88.5%, 91.2% for the plasma from colorectal cancer, lung cancer, and control plasma samples, respectively, with 5-fold cross-validation (**Figure 4-12c, Figure 4-13**). For the incorrectly classified samples, we noticed that 4 out of 5 colorectal cancer plasma were from metastatic colorectal cancer patients while the fifth was in fact tubular adenoma. In the case of lung cancer, one misclassified sample came from a patient with benign fibrous tissue. Taken together, we demonstrated for the first time that both tumor load and tissue of origin can be quantitatively characterized by methylation haplotype analysis of cell free DNA in plasma.

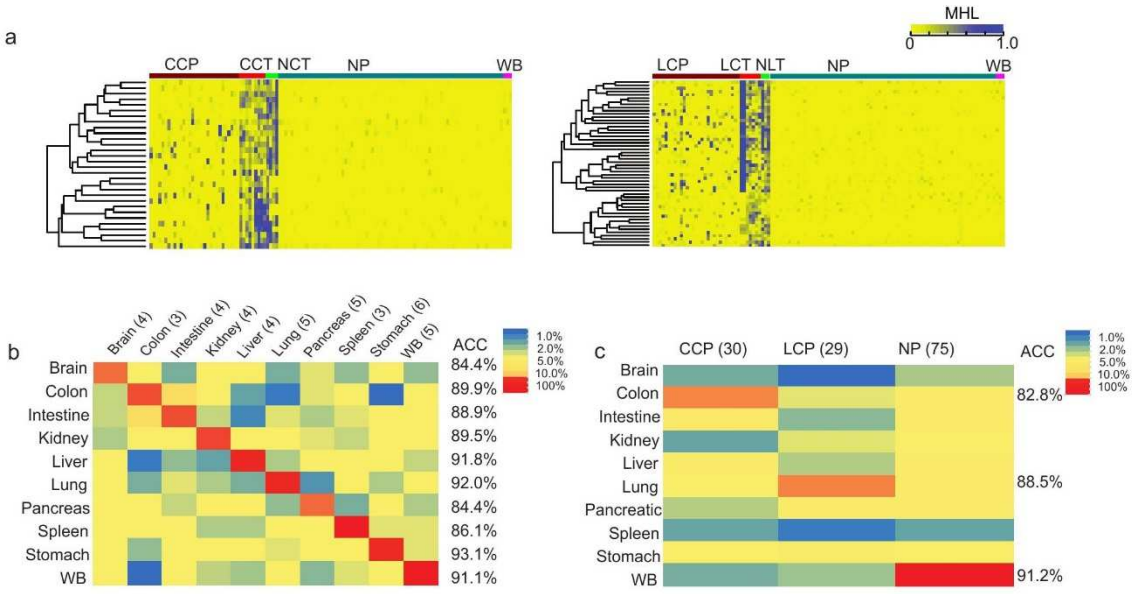


Figure 4-12. MHL-based prediction of cancer tissue of origin from plasma DNA
(a) Detection of tissue-specific MHL in the plasma of patients with cancer but not in plasma or whole blood from healthy individuals. Tissue-specific MHLs were visible in corresponding tissue and plasma samples from patients with cancer, indicating the feasibility for tissue-of-origin mapping. (b) Identification of informative MHBs for tissue prediction, using training data included in the WGBS and RRBS data sets from reference tissues, number of replicates shown in parentheses. The color key indicates the tissue-of-origin mapping accuracy (ACC). (c) Application of the prediction model to plasma samples from patients with colorectal cancer (n = 30) or lung cancer (n = 29) and from healthy individuals (n = 75)

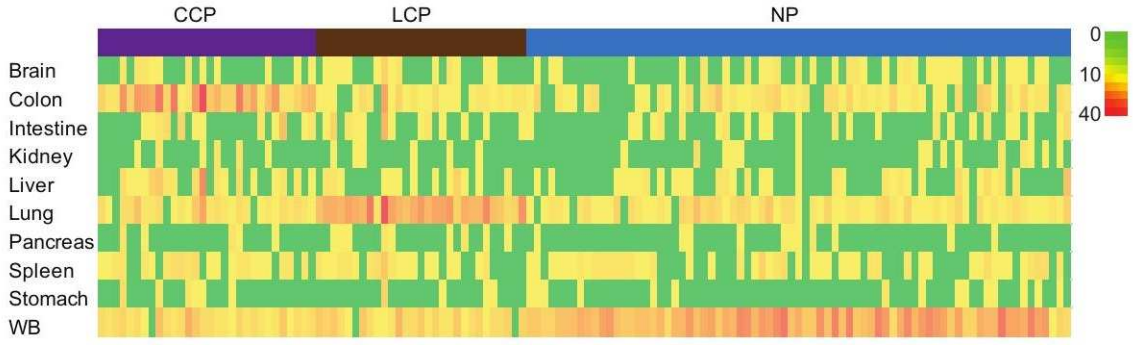


Figure 4-13. Distribution of tissue-specific MHBs counts in human plasma samples.
Color bar represents the number of tissue specific MHBs (for each respective tissue) over the MHL threshold in each plasma sample.

Discussion

Here we extended a well-established concept in population genetics, linkage disequilibrium, to the analysis of co-methylated CpG patterns. Although the mathematical representations are identical, there are two key differences. First, traditional linkage disequilibrium was defined on human individuals in a population, whereas in this study the analysis was performed on the diploid genome of individual cells in a heterogeneous cell population. Second, linkage disequilibrium in human populations depends on the mutation rate, frequency of meiotic recombination, effective population size and demographic history. The LD level decays typically over the range of hundreds of kilobases to megabases. In contrast, CpG co-methylation depends on DNA methyltransferases and demethylases, which tend to have much lower processivity (if any), and, in the case of hemi-methyltransferases, much lower fidelity than DNA polymerases[100]. Therefore, methylation LD decays over much shorter distance (in tens to hundreds of bases), with the exception of imprinting regions. Even if longer-read-sequencing methods were used, we did not expect a radical change of the block-like pattern presented in this work, which is supported by another recent study[86]. Nonetheless, these short and punctuated blocks capture discrete entities of epigenetic regulation in individual cells that are widespread in the human genome. This phenomenon can be harnessed to improve the robustness and sensitivity of DNA methylation analysis, such as the deconvolution of data from heterogeneous samples including cfDNA.

While we demonstrated a superior power of MHL over single-CpG methylation level or average methylation level in classification and deconvolution, the accuracy is slightly less than what has been reported on the deconvolution of blood cell types. One major difference is that each reference tissue type itself is a mixture of multiple cell types

that might share various degrees of similarity with another reference tissue type. Furthermore, most solid tissues also contain blood vessels and blood cells. Given such background signals, the accuracy that we achieved is very promising, and will be further improved once reference methylomes of pure adult cell types are available.

Practically, the amount of cell-free DNA per patient is rather limited, typically in the range of tens to hundreds of nanogram. We used 1 to 10 ng per patient for the sc-RRBS experiment. Considering the material losses during bisulfite conversation and library preparation, as well as the sequencing depth, there were most likely no more than 30 genome equivalents in each data set. Our data set is rather sparse, especially when the fraction of tumor DNA is low. Hence the chance of finding cancer-specific methylation haplotypes in a specific region consistently across many samples is low. This is likely the reason that marker sets selected based on random forest has limited sensitivity and specificity. However, epigenetic abnormalities tend to be more widespread across the genome (compared with somatic mutations), and hence we were able to integrate the sparse coverage across many loci to achieve very accurate prediction by direct counting of methylated haplotypes with the appropriate tissue-specific features. Further technical improvements on sample preparation and library construction, combined with larger sets of patient and normal plasma, will undoubtedly increase the coverage and further improve the specificity and sensitivity to the level required for clinical diagnosis.

Methods

Processing of human normal tissues

Ten human primary normal tissues were purchased from BioChain. Approximately 200 ng of genomic DNA from ten human primary tissues in the volume of 50 µL was fragmented into an average size of 400 bp in a Covaris micro TUBE with

Covaris E210 ultrasonicator. Fragmented genomic DNA was converted into Illumina paired-end sequencing libraries using KAPA Library Preparation kit (KAPA Biosystems) following manufacturer's instruction with modifications. After end-repair and dA-tailing, ligation with methylated adapters was performed at 20 ºC for 15 min in the presence of 10-fold molar excess of Illumina methylated adapters (Illumina). The ligation mixture was purified with an equal volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted with 23 µL of 10mM Tris-HCl, pH8.5. Next, 20 µL of adaptor ligated DNA was bisulfite converted using EZ DNA Methylation-Lightning kit (Zymo Research) following manufacturer's protocol and eluted with 30 µL of 10mM Tris-HCl, pH8.5. Bisulfite converted DNAs were amplified using iQ SYBR Green Supermix (Bio-Rad) with 200 nM each of PCR primer PE1.0 and multiplexing PCR primer for 10 cycles in 100 µL total volume. PCR products were purified with 0.8X volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted with 50 µL of 10mM Tris-HCl, pH8.5, pooled in equimolar ratios, and size selected using 6% TBE gels for 400-600 bp. The concentration of sequencing libraries was quantified by qPCR using KAPA Library Quantification kit (KAPA Biosystems). Libraries were sequenced on HiSeq2500 for PE 100 cycles.

Processing of patient tumor tissues.

Cancer tissue and plasma samples were collected from UCSD Moores Cancer Center. Clinical information, gender, age and TNM staging, on the patients was limited because the samples were de-identified. Informed consent was obtained from all subjects. All the samples are diagnosis to corresponding cancers according to the World Health Organization classification criteria. 88.4% samples were derived from Caucasian population while 6.8% and 3.3% samples were from Asian and African population (detail see **Table 4-8**). Genomic DNAs were extracted from 20-50 mg of primary tumor tissues from lung, colon and pancreatic cancer patients using DNeasy Blood and Tissue kit

(QIAGEN) following the manufacturer's instruction and eluted in 400 µL of AE buffer (QIAGEN). The concentration and quality of genomic DNA were assessed by Qubit dsDNA HS Assay kit (Life Technologies) and NanoDrop (Thermo Scientific), respectively. To generate RRBS sequencing libraries, 100 ng of gDNA were digested with 20 U of *Msp*I (Thermoscientific) in 1X Tango buffer (Thermoscientific) and 1 ng of unmethylated lambda DNA (Promega) in order to assess for bisulfite conversion rate in 30µL total volume for 3 h at 37 ºC and heat inactivated at 65 ºC for 20 min. Next, 5U of Klenow fragment, exo- (Thermoscientifc) and a mixture of dATP, dGTP, and dCTP (New England Biolabs) were added to *Msp*I-digested DNAs for a final concentration of 1 mM, 0.1 mM, and 0.1 mM for dATP, dGTP, and dCTP, respectively in 32 µL for end-repair and dA-tailing. The mixture was mixed and incubated at 30 ºC for 20 min, 37 ºC for 20 min, and heat inactivated at 75 ºC for 10 min. dA-tailed DNA was purified with 2X volume of Agencourt AMPure XP beads (Beckman Coulter) and resuspended dA-tailed DNA with 20 µL nuclease-free water without discarding the magnetic beads. dA-tailed DNAs were then ligated to methylated adaptors in 30 µL total volume containing 30 U of T4 DNA ligase, HC (Thermoscientific), 1X Ligation buffer (Thermoscientific), and 500 nM individual TruSeq multiplexing methylated adaptors (Illumina). The ligation mixture was mixed well and incubated at 16 ºC for 20 h, heat inactivated at 65 ºC for 20 min, purified by adding 60 µL of PEG 8000/5M NaCl buffer (Teknova) to adaptor ligated DNA and bead mixture, and eluted in 20 µL of nuclease-free water. Next, the adaptor ligated DNA were bisulfite converted using the MethylCode Bisulfite Conversion kit (Life Technologies) following manufacturer's protocol and eluted in 35 µL of Elution buffer (Life Technologies). Bisulfite treated DNAs were amplified using 5 U of PfuTurboCX (Agilent Technologies) and 300 nM each of TruS_F and TruS_R primers for 14 cycles in 100 µL total volume. PCR products were purified with an equal volume of Agencourt

AMPure XP beads (Beckman Coulter) and eluted with 50 µL of 10mM Tris-HCl, pH8.5, pooled in equimolar ratios, and size selected using 6% TBE gels for 150-400 bp. The concentration of sequencing libraries was quantified by qPCR using KAPA Library Quantification kit (KAPA Biosystems). Libraries were sequenced on Illumina HiSeq2500 for PE 100 cycles.

Processing of plasma samples

Normal plasma samples were obtained from UCSD Shirley Eye center. Information such as gender and age was limited because the samples were de-identified. Informed consent was obtained from all subjects. Plasma samples from patients were processed using the QIAamp Circulating Nucleic Acid Kit (Qiagen) to extract circulating DNA. The DNA extracted from plasma were then concentrated using ethanol precipitation and eluted in 15 µL nuclease-free water. Next, 1-10 ng of DNA were digested with 10 U of *Msp*I (Thermoscientific), 1X Tango buffer (Thermoscientific), and 10 pg of unmethylated lambda DNA (New England Biolabs) as control for ~13 h at 37 ºC, then heat inactivated at 65 ºC for 20 min. Next, 5 U of Klenow fragment, exo- (Thermoscientifc) and a mixture of dATP, dGTP, and dCTP (New England Biolabs) were added for a final concentration of 1 mM, 0.1 mM, and 0.1 mM for dATP, dGTP, and dCTP respectively. The mixture was gently vortexed, and incubated at 30 ºC for 20 min, 37 ºC for 20 min, and finally 75 ºC for 10 min. To perform adaptor ligation, the dA-tailed DNA were added to a 5 µL mixture of 1X Tango buffer, 30 U of T4 DNA Ligase, HC (Thermoscientific), 2.5 mM ATP, and 500 nM individual TruSeq multiplexing methylated adaptors. The combined mixture was gently vortexed, incubated at 16 ºC for ~20 h, then heat inactivated at 65 ºC for 20 min. The ligation mixture was purified using Agencourt AMPure XP beads (Beckman Coulter), and eluted in 20 µL of nuclease-free water. The ligated products were then bisulfite converted using the MethylCode Bisulfite Conversion

kit (Life Technologies). Two rounds of amplification were performed after bisulfite conversion. The first round was using PfuTurboCX (Agilent Technologies) for 12 cycles in 50 µL total volume, then the second round was performed using Phusion HotStart Flex (New England Biolabs) master mix for 9 cycles in 50 µL total volume. Final PCR products were purified, pooled in equimolar ratios, and size selected using polyacrylamide gels for 150-400 bp. Libraries were sequenced on both Illumina MiSeq and HiSeq2500 for PE 100 cycles.

NGS read mapping

WGBS and RRBS data were processed in similar fashions. We first trimmed all PE or SE fastq files using trim-galore version 0.3.3 to remove low quality bases and biased read positions. Next, the reads were encoded to map to a three-letter genome via conversion of all C to T or G to A if the read appears to be from the reverse complement strand. Then the reads were mapped using BWA mem version 0.7.5a, with the options "-B2 -c1000" to both the Watson and Crick converted genomes. The alignments with mapping quality scores of less than 5 were discarded and only reads with a higher best mapping quality score in either Watson or Crick were kept. Finally, the encoded read sequences were replaced by the original read sequences in the final BAM files. Overlapping pair end reads were also clipped with bamUtils clipOverlap function.

Identification of methylation haplotype blocks

Human genome was split into non-overlapping "sequenceable and mappable" segments using a set of in-house generated WGBS data from 10 tissues of a 25-year adult male donor. Mapped reads from WGBS data sets were converted into methylation haplotypes within each segment. Methylation linkage disequilibrium was calculated on the combined methylation haplotypes. We then partitioned each segment into

methylation haplotype blocks (MHBs). MHBs were defined as the genomic region in which the r^2 value of two adjacent CpG sites is no less than 0.5.

High methylation linkage regions defined based on ENCODE and TCGA data.

We collected RRBS data from the ENCODE project (downloaded from UCSC Genome Browser) and HM450K data from the TCGA project. Pearson correlation coefficient were calculated between adjacent CpG sites across all samples. The Takai and Jones's sliding-window algorithm[101] was used to identify blocks of highly correlated methylation. We set a 100-base window in the beginning of genomic position and move the window to the downstream when there are least 2 probes in the window. Calculate the total probes in extended regions until the last window does not meet the criteria. The regions covering at least 4 probes were defined as CpG dense regions, and the average Pearson correlation coefficients among all the probes in cancer and normal samples were calculated respectively. Simulation analysis to investigate the relationship between LD at the single-read level and correlation coefficients of average 5mC between two CpG sites were performed based on random sampling of 10 different methylation haplotypes from each of the 1000 individuals.

Enrichment analysis of methylation haplotype blocks for known functional elements

Enrichment analysis was performed by random sampling as previously described[102]. Genomic regions with same number (147,888), fragment length distribution and CpG ratios were randomly sampled within the mappeable regions (genomic regions beyond CRG mappability blacklisted regions and non-cover regions in our WGBS dataset), and repeated 1,000,000 times. Statistical significance was estimated based on empirical p-value (*P*). Fold changes (enrichment factors) were calculated as the ratios of observation over expectation. Exon, intron, 5-UTR, 3-UTR were collected UCSC

database. Enhancer definition was based on Andersson et al. [70], super enhancer was derived by Hnisz et al.[103] and promoter regions were based on the definition by Thurman et al.[71]. All the genomic coordinates were based on GRCh37/hg19.

Calculating methylation haplotype load

We defined a methylated haplotype load (MHL) for each candidate region, which is the normalized fraction of methylated haplotypes at different length:

Eq. 4-1 $$\mathrm{MHL} = \frac{\sum_{i=1}^{l} w_i \times P(MH_i)}{\sum_{i=1}^{l} w_i}$$

Eq. 4-2 $$w_i = i$$

Where l is the length of haplotypes, $P(MH_i)$ is the fraction of fully successive methylated CpGs with i loci. For a haplotype of length L, we considered all the sub-strings with length from 1 to L in this calculation. w_i is the weight for i-locus haplotype. Options for weights are $w_i = i$ or $w_i = i^2$ to favor the contribution of longer haplotypes. In the present study, $w_i = i$ was applied.

Following the concept of Shannon entropy $H(x)$, methylation entropy (ME) for haplotype variable in specific genome region were calculated with the following formula:

Eq. 4-3 $$H(x) = -\sum_{i=1}^{l} P(x) \times \log_2 P(x)$$

Eq. 4-4 $$\mathrm{ME} = -\frac{1}{b}\sum_{i=1}^{n} P(H_i) \times \log_2 P(H_i)$$

Eq. 4-5 $$P(H_i) = \frac{h_i}{N}$$

For a genome region with b CpG loci and n methylation haplotype, $P(H_i)$ represents the probability of observing methylation haplotype H_i, which can be calculated by dividing the number of reads carrying this haplotype by the total reads in this genomic region. ME is bounded between 0 and 1, and can be directly compared across different regions genome-wide and across multiple samples. Methylation entropy were widely used in the measurement of variability of DNA methylation in specific genome regions[104].

Epi-polymorphism[105] was calculated as:

Eq. 4-6 $$\mathrm{ppoly} = 1 - \sum_{i=1}^{n} P_i^2$$

where P_i is the frequency of epi-allele i the population (with 16 potential epialleles representing all possible methylation states of the set of four CpGs).

Developmental germ layers and tissue specific MHBs.

To investigate the germ layer and tissue specific MHBs, group specific index (GSI, see below) was defined. An empirical threshold GSI>0.6 was used define layer and tissue specific MHBs. Layer specific MHBs were selected again to show the ability to distinguish different development layers. Tissue specific MHBs were further used for tissue mapping and cancer diagnosis.

Eq. 4-7 $$\mathrm{GSI} = \frac{\sum_{j=1}^{n} 1 - \frac{\log_2(\mathrm{MHL}(j))}{\log_2(\mathrm{MHL}_{max})}}{n-1}$$

n indicates the number of the groups. $\mathrm{MHL}(j)$ denotes the average of MHL of j^{th} group. $\mathrm{MHL}(\max)$ denotes the average of MHL of highest methylated group.

Genome-wide methylation haplotype load matrix analysis

Methylation haplotype load was calculated for all MHBs on each sample. The MHBs with top 15% MHL were included in the heatmap to investigate the tissue relationship. The Euclidean distance and Ward.D aggregation were used in the heatmap plot. PCA was conducted with default setting of the corresponding R packages. Before the PCA analysis, raw data were quantile normalized within same tissue/cell groups. Standardization (scale) and batch effect elimination (the Combat algorithm[106]) were also applied to decrease the random noise. MAF and IMF were extracted from BAM files with customized PileOMeth (https://github.com/dpryan79/PileOMeth). Differential MHL analysis between cancer plasma and normal plasma were based on two-tailed Student's

t-test or Wilcoxon rank sum test. Correction for multiple testing was based on false discovery rate (FDR). Statistic variations were estimated among different groups and therefore one-way ANOVA analysis could be conducted.

Simulation and real-data deconvolution analysis

Deconvolution analysis was performed on simulated and experimental datasets. The deconvolution references were constructed on data from human normal primary tissues, whole blood (WB), colorectal cancer tissues (CCT) and lung cancer tissues (LCT). For the simulation analysis, methylation haplotypes from CCT and WB were randomly mixed to generate a series of synthetic data sets with CCT factions ranging from 0.1% to 50%. We then plotted the expected and observed CCT factions. Although MHL is a non-linear metrics, when mixing CCT and WB, we found the deconvolution result is accurate with log-transform (median root-mean-square-error < 5%), which is within the acceptable region of the deconvolution method[107] when the contribution of colorectal fraction is less than 20%. Tissue specific MHBs were selected features for deconvolution based on non-negative decomposition with quadratic programming[31,107,108]. MHL values were log-transformed before deconvolution.

Highly methylated haplotype in cancer plasma and normal tissues

Highly methylated haplotype (HMH) was defined as the methylation haplotype that have at least 2 methylated CpGs in the haplotype. Cancer-associated highly methylated haplotypes (caHMH) were the ones only found in cancer plasma samples but absence in any of the normal plasma samples and normal tissues. For the analysis of matched tumor-plasma data from the same individuals, caHMHs were the HMHs present in both the cancer plasma and the matched primary cancer tissues, but absence in all normal samples. In the analysis of plasma samples with no matched primary tumor

tissue, we identified caHMHs by subtracting HMHs found in cancer plasma with those present in all normal tissues and all normal plasma samples.

Simulation of MHL in plasma mixture and comparison between MHL and 5mC in the plasma mixture

In evaluating caHMHs as potential markers for non-invasive diagnosis, we hypothesized that cfDNA in plasma is a mixture of DNA fragments from cancer cells and WB cells at different ratios (cancer DNA fragment from 0.1% to 50%). We created synthetic mixtures by random sampling of haplotypes in the Group II regions from cancer and WB data sets at different ratios, and repeated 1,000 times to empirically determined the mean and variance of MHL and 5mC levels at different fractions of cancer DNA. Once an empirical “standard curve” was constructed, we then used it to estimate the fraction cancer DNA in the plasma samples. In addition, we assessed the relationship between estimated cfDNA fraction and log-transformed normalized plasma cfDNA yield by linear regression. Signal-to-noise ratio to MHL and 5mC was conducted with the 1,000-time sampling procedure and then the average estimated tumor fraction as well as the variation (standard deviation) were recorded and the ratio was calculated to measure the performance of the metric.

Cancer tissue-of-origin analysis with plasma DNA.

Tissue specific methylation haplotype blocks (tsMHBs) were identified by a 2-tailed t-test with FDR correction. Additional statistical analyses with MHL were also conducted by 2-tailed t-test unless stated explicitly. CRC plasma and LC plasma prediction evaluation were applied by random forecast therefore the test and validation sample were independent. Tissue-of-origin prediction was performed using a tsMHBs counting strategy, in which the tissue-of-origin of the plasma were assigned to the reference group with the maximum number of tsMHB fragments (assignment by

maximum likelihood). Specifically, in the first stage, the tissue-specific MHBs were identified with WGBS and RRBS datasets from solid tissues in the training samples. tsMHBs (each tissue have ~ 300 MHBs) were identified with the cutoff GSI> 0.1. In the second stage, the predictions were validated with our own RRBS dataset that included 30 colorectal cancer plasma, 29 lung cancer plasma and 75 normal plasma samples. In the test dataset, we separated the samples into 5 parts so that 5-fold cross-validation could be applied to estimate the stability of the prediction, and the number of tissue-specific MHB features were iterating from 50 to 300. The minimum number of features was selected when the accuracy for cancer plasma is higher than 0.8 and the accuracy for normal plasma is higher than 0.9 since we require high specificity in clinical applications. The selected number of features were used in the remaining samples to measure the accuracy of tissue-mapping. The variations of sensitivity, specificity, and accuracy in different subsets of 5-fold cross-variation were low (training dataset standard deviation<0.04 while testing dataset standard deviation<0.14）

Supplementary Tables

Table 4-7. WGBS datasets information and mapping statistics

Sample	Source	Tissue type	Tissue	Total mapped reads	Ave. depth of coverage	Genomic overage
N37-Cerebellum (CRBL)	This study	normal tissue	Cerebellum	96,004,220	3.81	85.82%
N37-Colon	This study	normal tissue	Colon	86,362,732	3.49	84.27%
N37-Frontal lobe(FL)	This study	normal tissue	Frontal lobe	73,138,777	3.06	81.55%
N37-Heart	This study	normal tissue	Heart	73,609,833	3.08	81.30%
N37-Small intestine(SI)	This study	normal tissue	Small intestine	84,071,507	3.41	83.91%
N37-Liver	This study	normal tissue	Liver	92,657,701	3.70	85.74%
N37-Lung	This study	normal tissue	Lung	89,779,805	3.61	85.03%
N37-Skeletal muscle(SM)	This study	normal tissue	Skeletal muscle	105,158,705	4.11	87.52%
N37-Pancreas	This study	normal tissue	Pancreas	108,799,699	4.39	84.28%
N37-Stomach	This study	normal tissue	Stomach	82,811,557	3.47	80.86%
methylC-seq_h1+bmp4_r1	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	508,320,946	14.43	92.69%
methylC-seq_h1+bmp4_r2	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	596,457,521	17.02	93.65%
methylC-seq_h1-msc_r1	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	544,860,203	19.74	92.13%
methylC-seq_h1-msc_r2	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	235,582,915	10.18	77.19%
methylC-seq_h1-npc_r1	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	717,247,821	19.55	93.53%
methylC-seq_h1-npc_r2	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	636,750,674	18.16	92.97%
methylC-seq_h1_mesendoderm_r1	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	586,720,277	20.23	94.10%
methylC-seq_h1_mesendoderm_r2	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	337,751,553	11.78	93.44%
methylC-seq_h1_r1	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	496,854,703	11.75	92.45%
methylC-seq_h1_r2	PMCID:PMC3786220	stem cells and progenitors	embryonic stem cells derived	548,343,316	16.09	93.96%
Centenarian	PMID:22689993	normal tissue	white blood cells	437,865,307	13.88	93.19%
Middle-age	PMID:22689993	normal tissue	white blood cells	436,803,766	13.89	93.10%
New-born	PMID:22689993	normal tissue	white blood cells	446,752,526	14.14	93.34%
Colon_primary_tumor	PMID:23925113	primary tumor tissue	primary colon tumor	930,844,352	30.26	88.49%
HCT116	PMID:25239471	cell line	colon cancer cell line	313,233,043	8.12	92.45%
STL001BL-01	Roadmap Epigenetics	normal tissue	Bladder	2,225,282,265	78.25	92.13%
STL001FT-01	Roadmap Epigenetics	normal tissue	Fat	807,909,068	28.38	91.51%
STL001GA-01	Roadmap Epigenetics	normal tissue	Gastric	866,472,335	30.44	91.62%
STL001LG-01	Roadmap Epigenetics	normal tissue	Lung	782,767,450	27.71	91.51%
STL001LV-01	Roadmap Epigenetics	normal tissue	Heart	1,824,013,758	64.39	92.00%
STL001PO-01	Roadmap Epigenetics	normal tissue	Muscle	756,086,418	26.53	91.46%
STL001RV-01	Roadmap Epigenetics	normal tissue	Heart	708,173,343	25.18	91.38%
STL001SB-01	Roadmap Epigenetics	normal tissue	Intestine	1,987,847,032	69.53	92.15%
STL001SG-01	Roadmap Epigenetics	normal tissue	Colon	2,119,821,939	74.72	92.13%
STL001SX-01	Roadmap Epigenetics	normal tissue	Spleen	1,041,390,267	36.66	91.69%
STL001TH-01	Roadmap Epigenetics	normal tissue	Thymus	1,861,688,106	65.81	92.02%
STL002AD-01	Roadmap Epigenetics	normal tissue	Kidney	951,581,565	33.76	91.17%
STL002AO-01	Roadmap Epigenetics	normal tissue	Vessel	697,441,452	24.63	90.97%
STL002EG-01	Roadmap Epigenetics	normal tissue	Esophagus	868,051,856	30.79	91.53%
STL002FT-01	Roadmap Epigenetics	normal tissue	Fat	1,004,247,873	35.60	91.20%
STL002GA-01	Roadmap Epigenetics	normal tissue	Gastric	762,365,245	26.91	91.04%
STL002LG-01	Roadmap Epigenetics	normal tissue	Lung	2,115,617,666	74.69	91.70%
STL002OV-01	Roadmap Epigenetics	normal tissue	Ovary	2,129,375,638	75.65	91.59%
STL002PA-01	Roadmap Epigenetics	normal tissue	Pancreas	844,006,942	29.73	91.11%
STL002PO-01	Roadmap Epigenetics	normal tissue	Muscle	866,606,093	30.54	91.14%
STL002SB-01	Roadmap Epigenetics	normal tissue	Intestine	563,351,692	21.55	84.93%
STL002SX-01	Roadmap Epigenetics	normal tissue	Spleen	953,725,789	33.63	91.21%
STL003AD-01	Roadmap Epigenetics	normal tissue	Kidney	2,033,812,621	71.71	92.03%
STL003AO-01	Roadmap Epigenetics	normal tissue	Vessel	3,216,930,660	112.34	92.32%
STL003EG-01	Roadmap Epigenetics	normal tissue	Esophagus	2,317,477,506	81.45	92.18%
STL003FT-01	Roadmap Epigenetics	normal tissue	Fat	1,866,543,974	65.41	92.03%
STL003GA-01	Roadmap Epigenetics	normal tissue	Gastric	2,225,019,266	77.85	92.11%
STL003LV-01	Roadmap Epigenetics	normal tissue	Heart	1,990,945,522	70.05	92.08%
STL003PA-01	Roadmap Epigenetics	normal tissue	Pancreas	1,762,389,162	62.11	92.04%
STL003PO-01	Roadmap Epigenetics	normal tissue	Muscle	2,296,688,359	81.23	92.09%
STL003RA-01	Roadmap Epigenetics	normal tissue	Heart	2,185,892,032	77.39	92.09%
STL003RV-01	Roadmap Epigenetics	normal tissue	Heart	2,056,445,884	72.16	92.09%
STL003SB-01	Roadmap Epigenetics	normal tissue	Intestine	790,000,067	28.13	91.47%
STL003SG-01	Roadmap Epigenetics	normal tissue	Colon	1,833,299,005	65.18	91.97%
STL003SX-01	Roadmap Epigenetics	normal tissue	Spleen	2,029,659,484	72.04	91.09%
STL011LI-01	Roadmap Epigenetics	normal tissue	Liver	1,101,355,323	39.08	91.76%
SRX381569_tumor_colon	PMID:26813288	primary tumor tissue	CRC tumor	1,046,795,688	32.55	86.00%
SRX381716_adenocarcinoma_lung	PMID:26813288	cancer cell line	H1437	530,779,493	16.99	85.88%
SRX381719_squamous_cell_tumor_lung	PMID:26813288	cancer cell line	H157	502,645,425	16.31	83.24%
SRX381722_small_cell_tumor_lung	PMID:26813288	cancer cell line	H1672	548,938,720	15.83	90.10%

Table 4-8. ENCODE RRBS dataset information

Sample ID	Project	Tissue of origin prediction training	Tissue type	Tissue
ENCFF000LUQ	ENCODE Project	Yes	normal tissue	Brain
ENCFF000LUU	ENCODE Project	Yes	normal tissue	Brain
ENCFF000LVA	ENCODE Project	Yes	normal tissue	Kidney
ENCFF000LVB	ENCODE Project	Yes	normal tissue	Kidney
ENCFF000LVJ	ENCODE Project	Yes	normal tissue	Liver
ENCFF000LVN	ENCODE Project	Yes	normal tissue	Liver
ENCFF000LVO	ENCODE Project	Yes	normal tissue	Lung
ENCFF000LVR	ENCODE Project	Yes	normal tissue	Lung
ENCFF000LVU	ENCODE Project	Yes	normal tissue	Pancreas
ENCFF000LVW	ENCODE Project	Yes	normal tissue	Pancreas
ENCFF000LWS	ENCODE Project	Yes	normal tissue	Stomach
ENCFF000LWW	ENCODE Project	Yes	normal tissue	Stomach
ENCFF000LVI	ENCODE Project	Yes	normal tissue	White blood cells
ENCFF000LVK	ENCODE Project	Yes	normal tissue	White blood cells

Table 4-9. Clinical characteristics of cancer patient samples

ID	Tissue	Description	Race
CRC.001	Colon	poorly differentiated adenocarcinoma, consistent with recurrent colonic adenocarcinoma	Caucasian
CRC.002	Colon	metastatic malignant hemangiopericytoma/malignant solitary fibrous tumor	Caucasian
CRC.003	Colon	sigmoid colon, rectum, and anus, resection. recurrent invasive moderately- to poorly-differentiated	Caucasian
CRC.004	Colon	metastatic adenocarcinoma.	Caucasian
CRC.005	Colon	metastatic mucinous adenocarcinoma.	Caucasian
CRC.006	Colon	metastatic moderately-differentiated adenocarcinoma with mucinous features.	Caucasian
CRC.007	Colon	adenocarcinoma	Caucasian
CRC.008	Colon	metastatic moderately-differentiated adenocarcinoma with munico features	Caucasian
CRC.009	Colon	adenocarcinoma with munico features, moderately differentiated, pt4an2b	Caucasian
CRC.010	Colon	mucinous adenocarcinoma, moderately differentiated	Caucasian
CRC.011	Colon	metastatic adenocarcinoma/primary-colon adenocarcinoma	Caucasian
CRC.012	Colon	invasive adenocarcinoma, moderately differentiated, two synchronous primary tumors, pt3n1c	Black
CRC.013	Colon	colonic mucosa with no diagnostic alteration.-negative for dysplasia or malignancy	Caucasian
CRC.014	Colon	metastatic adenocarcinoma with mucin production	Asian
CRC.015	Colon	colonic mucosa with ulcer, granulation tissue, and polarizable foreignmaterial with associated iron, see comment.-no malignancy identified	Caucasian
CRC.016	Colon	metastatic moderately differentiated adenocarcinoma, consistent withcolonic origin.	Caucasian
CRC.017	Colon	moderately differentiated colorectal adenocarcinoma	Caucasian
CRC.018	Colon	adenocarcinoma with mucinous features, moderately differentiated, pt4an2b	Caucasian
CRC.019	Colon	metastatic adenocarcinoma	Caucasian
CRC.020	Colon	Recurrent metastatic colon cancer	Caucasian
CRC.021	Colon	moderately-differentiated adenocarcinoma	Caucasian
CRC.022	Colon	tubular adenoma	Caucasian
CRC.023	Colon	moderately differentiated adenocarcinoma (rectosigmoid)	Caucasian
CRC.024	Colon	residual/recurrent adenocarcinoma, morphologically consistent with colorectal primar(abdominal wall)	Caucasian
CRC.025	Colon	residual invasive moderately differentiated adenocarcinoma	Caucasian
CRC.026	Colon	invasive moderately differentiated colonic adenocarcinoma	Caucasian
CRC.027	Colon	focal invasive adenocarcinoma, arising in a background of tubular adenoma with high-grade dysplasia	Caucasian
CRC.028	Colon	metastatic moderately differentiated adenocarcinoma consistent with colorectal origin	Caucasian
CRC.029	Colon	Metastatic rectosigmoid colon cancer to the liver	Caucasian
CRC.030	Colon	metastatic adenocarcinoma	Caucasian
LC.001	Lung	invasive moderately differentiated adenocarcinoma, pt26n0	Asian
LC.002	Lung	poorly differentiated squamous cell carcinoma. consistent with pt2n0mx, stage ib	Caucasian
LC.003	Lung	Carcinoid tumor of lung	Caucasian
LC.004	Lung	squamous cell carcinoma, clear cell variant, moderately differentiated	Caucasian
LC.005	Lung	squamous cell carcinoma, pt1bn0	Caucasian
LC.006	Lung	adenocarcinoma, mix subtype, including 60% bronchioalveolar carcinoma, mucinous subtype with 40% invasive carcinoma, acinar and papillary subtypes	Caucasian
LC.007	Lung	invasive squamous carcinoma, moderately differentiated, consistent with pt1n0mx	Caucasian
LC.008	Lung	invasive adenocarcinoma, poorly differentiated	Caucasian
LC.009	Lung	non-small-cell carcinoma	Caucasian
LC.010	Lung	poorly differentiated carcinoma, possibly adenoscc	Caucasian
LC.011	Lung	squamous cell carcinoma poorly differentiated, pt3n0	Caucasian
LC.012	Lung	adenocarcinoma	Caucasian
LC.013	Lung	well-differentiated neuroendocrine tumor arising in a background of neuroendocrine hyperplasia + tumorlets	Caucasian
LC.014	Lung	invasive adenocarcinoma, moderately to poorly differentiated, pt2an0	Caucasian
LC.015	Lung	cryptococcal pulmonary infection	Caucasian
LC.016	Lung	adenocarcinoma	Black or African American
LC.017	Lung	adenocarcinoma, acinar predominant, moderately differentiated	Caucasian
LC.018	Lung	adenocarcinoma	Caucasian
LC.019	Lung	adenocarcinoma, moderately differentiated	Caucasian
LC.020	Lung	moderately differentiated non-small cell adenocarcinoma	Asian
LC.021	Lung	squamous cell carcinoma	Caucasian
LC.022	Lung	non-small cell lung carcinoma; stage iiia	Caucasian
LC.023	Lung	adenocarcinmoa	Caucasian
LC.025	Lung	adenocarcinoma, mixed acinar and solid types, moderately differentiated	Asian
LC.026	Lung	non-small cell carcinoma	Caucasian
LC.027	Lung	adenocarcinom lung	Caucasian
LC.028	Lung	adenocarcinoma, poorly differentiated, pt2n0mx, stage 1b	Caucasian
LC.029	Lung	lung cancer (egfr mutation e746-a750 dels mutation in exon 19)	Black or African American
LC.030	Lung	lung adenocarcinoma.	Caucasian

Acknowledgements

We thank S. Kaushal for managing and handling patient samples in the UCSD Moores Cancer Center Biorepository Tissue Technology Shared Resource and S.M. Lippman, R. Liu and B. Ren for insightful discussions. This study was supported by US National Institutes of Health grants R01GM097253 (Kun Z.), R01 CA217642(Kang.Z), RO1EY025090 (Kang.Z.) and VA Merit Award (Kang.Z.) and P30CA23100 (S.M.L.)

Chapter 4 contains material as it appears in: Guo, Shicheng*, Dinh Diep*, Nongluk Plongthongkum, Ho Lim Fung, Kang Zhang, Kun Zhang. “Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA”. Nature Genetics. Used with permission. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 5: The development of methylation haplotype blocks in tumorigenesis

Introduction

Previous works in characterizing DNA methylation in cancer samples have discovered higher intrasample variability amongst cancer samples of the sample types[57]. In a recent study on 104 patients with primary chronic lymphocytic leukemias, local disordered DNA methylation[78] was discovered and higher disorder in promoters was further associated with decreased survival in these patients. These studies supports what have been suggested that there could be very few positively selected methylation changes in tumorigenesis and the general stochasticity of cancer methylome is a feature of cancer being associated with a plastic and stemlike epigenetic state[109] . From a biomarker development perspective then, most differentially methylated regions found between cancer and normal tissues are too stochastic to be good biomarker candidates.

In the previous chapter, we identified cancer associated highly methylated haplotypes belonging in methylation haplotype block regions as potential biomarkers for cancer detection in cell free DNA. This was not surprising since many studies have demonstrated that regions of frequent hypermethylation can be identified in cancer samples[110]. A recent study showed that loss of hydroxymethylation and subsequent gain of methylation in the gene bodies characterizes kidney tumorigenesis[11]. We hypothesized that cancer development could also be characterized by formation of methylation haplotype blocks within regions of hypermethylation in cancer and loss of hydroxymethylation could be one mechanism associated with the formation of these blocks.

In this chapter, we extended MHB identification to 107 WGBS datasets and 19 TAB-seq datasets. Next, we performed integrated analysis with arrays datasets to validate the MHBs followed by integrated analysis of multiple types of datasets to demonstrate the enrichment of MHBs at frequently hypermethylated regions in different

types of cancer, 5hmC marked regions in primary colon cancer tissues[111] and TET2 bound regions in HCT116 (colon cancer cell line)[111] and TET3 bound regions in HEK393T (human embryonic kidney cell line)[112]. Finally, we associated the MHB regions with DMRs of kidney tumorigenesis.

Results

Identification of an expanded set of MHBs from 107 WGBS

We collected 107 WGBS datasets, 61 of these were previous used in MHB identification (Chapter 4), and 46 additional datasets were as follows: 22 samples representing cancer cells and normal tissues from Heyn et al. 2016[89], 5 additional brain tissue and normal colon samples from Ziller et al. 2013[74], 4 primary normal and cancer kidney tissue samples from Chen et al. 2016[11], 15 normal blood subtypes and fetal tissues from the Roadmap Epigenomics project[84]. We applied the MHB finding algorithm to the combined dataset with ~20.9 billion haplotype informative reads. Unlike previously (Chapter 4), this set utilized haplotype reads with minimum 2 CpGs while previously a minimum of 4 CpGs was required for haplotype reads. We then applied a looser threshold of 0.3 linkage disequilibrium r^2 for MHBs (instead of 0.5 previously) due to the greater heterogeneity in this dataset and this resulted in 295,772 total MHBs. The expanded set of MHBs overlapped with 48% of the previous 147,888 MHBs. The loss of a majority of previous MHBs were probably due to increased heterogeneity (from ~0.711 to ~ 20.9 billion usable reads) or due to additional signals from 5hmCs in brain tissues (we added 8 brain tissue samples). The average methylation levels within these blocks are within 0 to 1, with bimodal peaks at 0.1 and 0.9 (**Figure 5-1a**). The average linkage disequilibrium was not less than 0.3 and has a peak around 0.4 (r^2) (**Figure 5-1b**). The fraction of discordant read was low, with a peak at 0.2 (**Figure 5-1c**).

To validate these expanded MHBs, we subset to CpGs from the Infinium HM450K assay. The adjacent CpG pairs within MHBs is significantly higher in linkage disequilibrium than the adjacent CpG pairs outside of MHBs (**Figure 5-1d**). Strong correlation between adjacent CpGs is necessary for CpGs pairs within MHBs. As expected, the pairwise Pearson's correlation for adjacent CpGs across 30 glioblastoma[10] and 4 cerebellum[41] samples (oxBS-array and BS-array dataset) showed significantly higher correlation for CpGs within MHBs than outside for both 5mC and 5hmCs (**Figure 5-1e,f**).

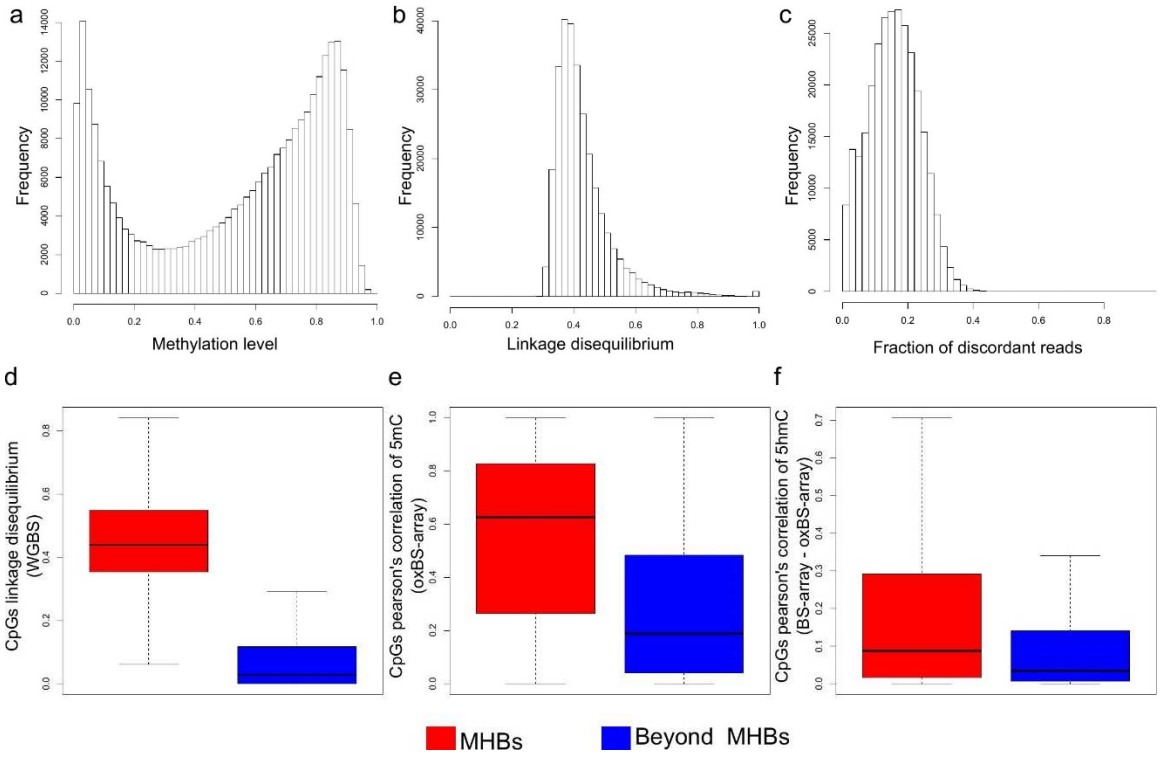


Figure 5-1. Expanded methylation haplotype blocks
(a) The distribution of average methylation across blocks from 107 WGBS samples. (b) the distribution of average linkage disequilibrium (r^2) values across blocks. (c) the fraction of discordant reads (reads with both methylated and unmethylated CpGs) within blocks. (d) the linkage disequilibrium (r^2) values for adjacent CpGs within blocks and beyond blocks. (e) the Pearson's correlation for methylation levels of adjacent CpGs within blocks and beyond blocks from oxBS-array data. (f) the squared Pearson's correlation for hydroxymethylation levels of adjacent CpGs within blocks and beyond blocks from oxBS-array subtracted from BS-array levels.

Extension of MHBs identification to TAB-seq datasets

We queried TAB-seq datasets for hydroxymethylated haplotype blocks using the same approach as the one used for WGBS datasets. We compiled a total of 19 TAB-seq datasets comprise of 5 samples represent stem and progenitor cell types (UCSD Human Reference Epigenome Mapping Project), 8 are blood subtypes[113,114], 2 are frontal cortex[115] samples, 2 primary kidney tissue[11] samples, and 2 primary kidney tumor tissue[11] samples. In total we obtained 3.4 billion haplotype informative reads and the MHB finding algorithm detected 27,422 hydroxymethylated haplotype blocks (hMHBs). The average hydroxymethylation level within these blocks has a peak around 0.05, which was less than the average hydroxymethylation level genome wide (**Figure 5-2a**). The average linkage disequilibrium within the blocks was not less than 0.3, and has a peak near 0.4 (r^2) (**Figure 5-2b**). The fraction of discordant reads was also very low, with a peak less than 0.05 (**Figure 5-2c**).

Again to validate these blocks, we subset to CpGs from the Infinium HM450K assay. The adjacent CpG pairs within hMHBs is significantly higher in linkage disequilibrium than the adjacent CpG pairs outside of MHBs (**Figure 5-2d**). The pairwise Pearson's correlation coefficient for adjacent CpGs across showed lower correlation for CpGs within hMHBs than outside for 5hmCs from a TAB-array dataset with human IPSCs, cardiovascular progenitor cells, neuroprogenitor cells, and human dermal fibroblasts (**Figure 5-2e**). We looked additionally at the set of 30 glioblastoma and 4 cerebellum oxBS-array datasets, and again did not see higher correlation for CpGs within hMHBs than outside hMHBs (**Figure 5-2f**). It was possible that 5hmC levels were too low within hMHBs, perhaps because in these regions there is stronger pressure to proceed to demethylation. Furthermore, we found that only 0.2% of CpGs covered by the Infinium HM450K arrays were within hMHBs while 11.7% of CpGs were

within MHBs, which make it difficult to validate hMHBs with pairwise correlation coefficient.

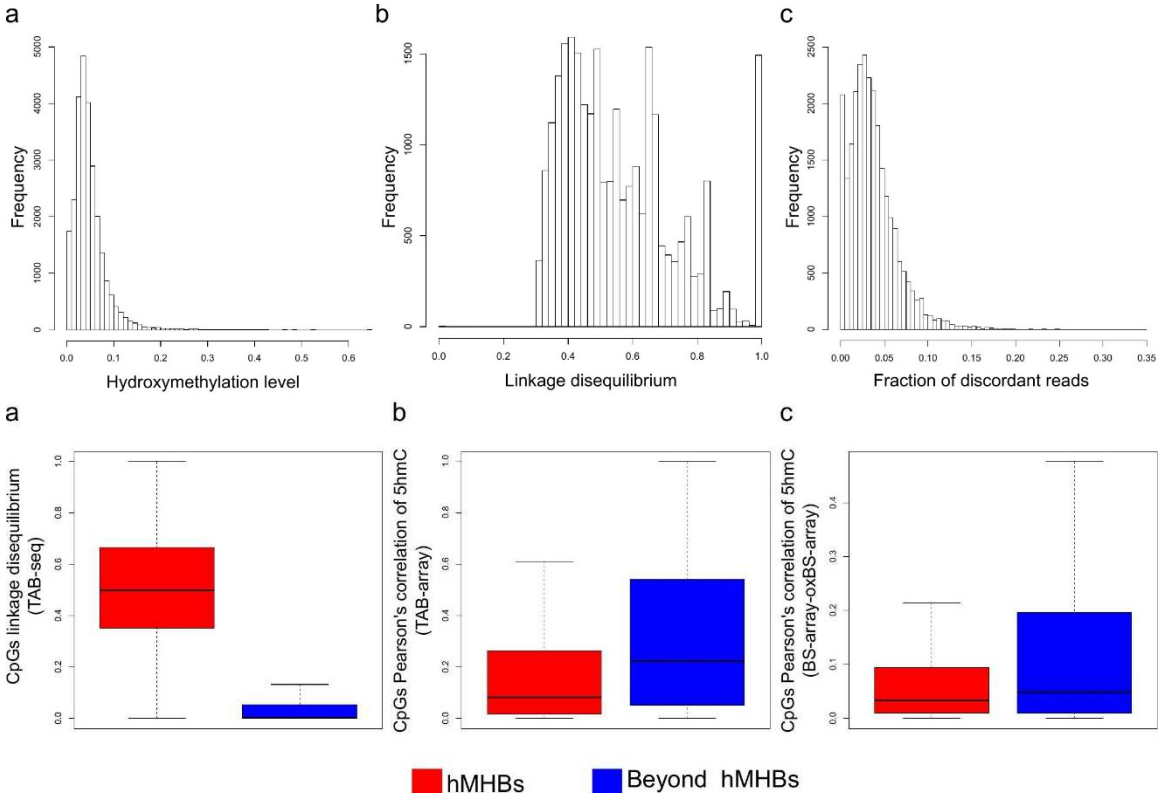


Figure 5-2. Hydroxymethylation haplotype blocks
(a) The distribution of average hydroxymethylation across blocks from 19 TAB-seq samples. (b) the distribution of average linkage disequilibrium (r^2) values across blocks. (c) the fraction of discordant reads (reads with both hydroxymethylated and unhydroxymethylated CpGs) within blocks. (d) the linkage disequilibrium (r^2) values for adjacent CpGs within blocks and beyond blocks. (e) the Pearson's correlation for hydroxymethylation levels of adjacent CpGs within blocks and beyond blocks from TAB-array data. (f) the squared Pearson's correlation for hydroxymethylation levels of adjacent CpGs within blocks and beyond blocks from oxBS-array subtracted from BS-array levels.

Regional enrichment analysis

Similar to our previous results (**Figure 5-3**), the expanded MHBs are enriched for imprinted genes, variably methylated regions (VMRs), CpG islands, enhancers, promoters, CpG island shores, and super-enhancers. They were also depleted at CpG island shelves, lamin-associated domains[91] (LADs), and large organized chromatin Lys9 modifications[92] (LOCKs). Not much unlike MHBs, the hMHBs were also enriched at

imprinted genes, VMRs, CpG islands, promoters, super-enhancers, and LADs. But they were not enriched for CpG island shores, shelves, and LOCKs.

Next, we asked whether the DMRs associated with cancer developments are associated with MHBs or hMHBs. First, we obtained a list of 1,154 CpG positions that have been identified as frequently hypermethylated in cancer from seven tissue types[110]. We performed regional enrichment analysis on this set and found a significant enrichment of 476 CpGs (permutation p-value =0, enrichment factor = 25.72) for MHBs, and a not significant enrichment of 5 CpGs (permutation p-value = 0.058, enrichment factor = 2.5) for hMHBs. We also obtained a merged set of DNA methylation valleys (DMVs)[13] which were identified from stem and progenitor cells as regions depleted of DNA methylation and commonly co-localizes with early developmental genes and genes involved in cancer pathways. Both MHBs and hMHBs were enriched within these DMVs, (permutation p-value=0, enrichment factor=12.12; permutation p-value=0, enrichment factor=2.84 for MHBs and hMHBs respectively). Since these region sets were defined based on analysis of DNA methylation only, it was not unexpected that hMHBs observed much less enrichment. However, the strong enrichments of our set of MHBs support the potential of MHBs as cancer biomarkers.

We also queried hmeDIP data for association with 5hmC peaks in normal colon and cancer colon tissues[111]. MHBs were significant enriched in 5hmC peaks (permutation p-value=0, enrichment factor=2.60; permutation p-value=0, enrichment factor=2.79 for normal and tumor colon respectively). However, hMHBs were depleted in 5hmC peaks likely because hMHBs have very low 5hmC levels that could not be easily detected by hmeDIP assay (permutation p-value=0, enrichment factor=0.94; permutation p-value=0, enrichment factor=0.84 for normal and tumor colon respectively).

We also looked for enrichment at TET2 binding peaks for colon cancer cells[111] (HCT116) and TET3 binding peaks for human embryonic kidney cells[112] (HEK293T). We found significant enrichment of peaks in MHBs for both TET2 and TET3 (permutation p-value =0, enrichment factor=3.55 and p=0, enrichment factor=4.96 for HCT116 and HEK293T respectively). There was just as strong enrichment of TET proteins binding for hMHBs in both cell types (permutation p-value =0, enrichment factor=4.39 and permutation p-value =0, enrichment factor=2.52 for HCT116 and HEK293T respectively).Thus, it appears that some regions with MHBs formation may associate with both 5hmC levels and Tet protein activity.

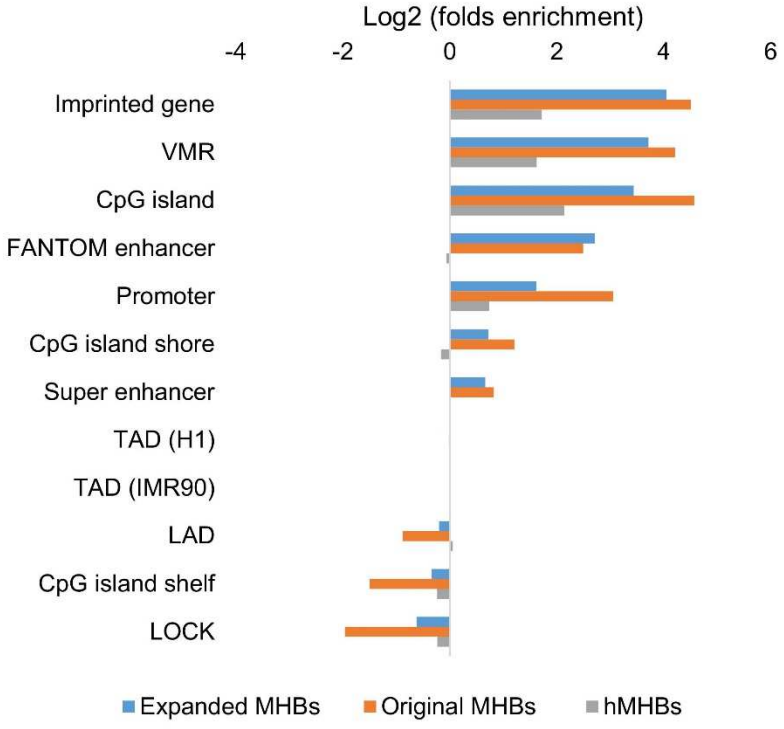


Figure 5-3. Regional enrichment of blocks

Association of 5hmC loss with MHBs

With the 295,772 MHBs as a starting point, we asked whether the DMRs associated with kidney tumorigenesis are associated with MHBs. We utilized TAB-seq and WGBS datasets from a kidney cancer study[11] comprised of two paired normal and tumor tissues samples to identify regions with loss of 5hmC and gain of 5mC. We identified 95,402 such DMRs and 26,334 of these DMRs overlaps with an MHB. Thus, these DMRs have significant overlap of MHBs (permutation p-value=0, enrichment factor=1.28). The kidney cancer study[11] also showed that downregulation of IDH1, which is a protein that typically generate the 2-ketoglutarate (2-KG) molecules utilized as a substrate for TET protein activities, is a mechanism underlying the loss of 5hmC in kidney cancer. In experiments with ectopic overexpression of IDH1, the study showed that the level of 5hmC can be restored in kidney cancer cells while also blocking tumor invasion in mice xenografts. Therefore a subset of MHBs are a result of lack of Tet2 activity that occurs normally in non-cancer tissues.

Conclusion

In this chapter, we investigated the relationship between methylation and hydroxymethylation in the context of linked methylation or haplotype blocks. Through an integrated analysis, we found support of MHBs as cancer biomarkers. We further discovered an association of 5hmC with MHBs and discovered a set of MHBs that co-localizes with DMRs in kidney tumorigenesis.

Methods

NGS read mapping

WGBS and TAB-seq data were processed in similar fashions. We first trimmed all PE or SE fastq files using trim-galore version 0.3.3 to remove low quality bases and biased read positions. Next, the reads were encoded to map to a three-letter genome via conversion of all C to T or G to A if the read appears to be from the reverse complement

strand. Then the reads were mapped using BWA mem version 0.7.5a, with the options "-B2 -c1000" to both the Watson and Crick converted genomes. The alignments with mapping quality scores of less than 5 were discarded and only reads with a higher best mapping quality score in either Watson or Crick were kept. Finally, the encoded read sequences were replaced by the original read sequences in the final BAM files. Overlapping pair end reads were also clipped with *bamUtils* clipOverlap function.

Identification of methylation haplotype blocks (MHBs) from WGBS or TAB-seq data

Human genome was split into non-overlapping "sequenceable and mappable" segments using a set of in-house generated WGBS data from 10 tissues of a 25-year adult male donor (same regions as Chapter 4). Mapped reads from WGBS data sets were converted into methylation haplotypes at minimum 2 CpG positions in Hg19. Methylation linkage disequilibrium was calculated on the combined methylation haplotypes for pairs of CpGs using equation 5.1 where F_{ij} is the fraction of total haplotypes with methylation patterns *i* and *j* for a pair of CpGs. Possible values for *i* (first position) and *j* (second position) are 'either' (X), 'unmethylated' (U) or 'methylated' (M). We then partitioned each segment into methylation haplotype blocks (MHBs). MHBs were defined as the genomic region in which the r^2 value of two adjacent CpG sites is no less than 0.3 and minimum 3 CpGs.

$$r^2 = \frac{(F_{MM} - F_{XM}F_{MX})^2}{F_{MX}F_{XM}F_{UX}F_{XU}}$$

Eq. 5-1

Identification of DMRs in kidney cancer

We performed DHMRs identification of kidney cancer TAB-seq datasets using *cgDMR-miner* (https://github.com/dinhdiep/cgDMR-miner). The TAB-seq dataset was comprise of two paired normal and tumor tissue samples. After DHMRs were identified using *cgDMR-miner*, we calculated the corrected average methylation frequencies for

each segment with BS-seq data from the same two paired normal and tumor tissue samples. The corrected average 5hmC level was calculated using equation 5-2, and the corrected average 5mC level was calculated using subtraction of the corrected 5hmC level from the BS-seq frequency. We then generated DHMRs with a minimum 10% loss of 5hmC for both pairs from normal to tumor. We then further required that the same region observe a 10% gain of 5mC from normal to tumor in both pairs. This resulted in 95,402 DMR regions.

Eq. 5-2 $$hmC = \frac{m_{tab}}{protection\ rate} - m_{wgbs}(1 - oxidation\ rate)$$

Enrichment analysis of methylation haplotype blocks for known functional elements

Enrichment analysis was performed by random sampling as previously described[102]. Genomic regions with same number and fragment length distribution were randomly sampled within the mappable regions (regions in our WGBS dataset), and repeated 1,000 times. Statistical significance was estimated based on the number of times an equivalent or higher number of overlapping regions were found. Fold changes (enrichment factors) were calculated as the ratios of observation over random expectation. Enhancer definition was based on Andersson et al.[70], super enhancer was derived from Hnisz et al.[103] and promoter regions were based on the definition by Thurman et al.[71]. All the genomic coordinates were based on GRCh37/hg19.

Chapter 6: Discussion and future directions

Through an iterative process of utilizing bioinformatics tools to develop better assays and utilizing experience with assays development to develop better bioinformatics tools, we demonstrated that the integration of bioinformatics with experimental biology can greatly advance research in biomarker development.

We first utilized bioinformatics to identify informative genomic regions that could be considered “hotspots” for identifying differentially methylated regions and developed a scalable and high throughput method to assay to investigate these regions. Leveraging our assay development expertise, we identified the areas in which methylation sequencing analysis pipeline could be improved. This led to the development of a highly accurate pipeline with an emphasis towards better accuracy that no other tools can provide. Specifically, we implemented reads trimming to remove sequencing bias, adaptor trimming, an alignment approach for longer read lengths with higher specificity, and SNP calling on bisulfite reads that helps with sample identification.

We demonstrated that BSPP is a flexible framework for developing screening panels. For example, panels of “hotspots” probes could be designed for the most informative CpGs in each disease context. Additionally, two technical advantages for utilizing BSPP over other targeted approaches are the ability to perform absolute molecule counting with the capture probes and higher mappability of the data since only regions with unique reads mapping are selected for capture. One immediate area of interest for panel development is for the quantification of 5hmC which are found at only a fraction of the 5mC levels.

We next asked whether novel tools for identifying potential DNA biomarkers can be created with more sensitivity for making discoveries from reference data. This led us to develop *cgDMR-miner*, a bioinformatics tool for identifying differentially methylated regions (DMRs) with much better sensitivity on shallow sequencing data than current

methods. Another novelty in our method is the ability identify DMRs from datasets with multiple groups that can include ungrouped samples. From the biomarkers development perspective, it is cost effective to be able to identify potential biomarker candidates with shallow sequencing datasets in initial screens and validate biomarkers in a larger screen with a targeted assay.

We have only tested *cgDMR-miner* on bulk libraries but not on extremely low input libraries. Applicability of *cgDMR-miner* to sorted, 10-30 cells datasets should also be investigated. Low input libraries are susceptible to higher levels of PCR bias, thus, questions regarding how tolerance our approach is to various levels of technical noise in data such as batch effects and PCR bias need to be answered in future works.

Lastly, we were also concerned with current limitations in investigating DNA methylation when the DNA sample is in small amount such as cell free DNA. Before, only targeted sequencing or whole genome bisulfite sequencing were successfully applied to study DNA methylation in cell free DNA[30,31,116]. We were the first group to apply sc-RRBS to cell free DNA and achieved an apparent enrichment of RRBS targets of about 2 folds. To take advantage of this enrichment of CpG dense regions, we developed a bioinformatics approach focused on regions with coordinated methylation of nearby CpGs. This approach relies on the identification of methylation haplotype blocks, or CpG regions with highly linked methylation and relies on the application of the methylation haplotype load (MHL) metric to capture the pattern of co-methylation on single DNA molecules. With this, we were able to identify hypermethylated regions in cancer and detect methylated haplotypes (physical molecules) from these regions in patients cell free DNA.

Future investigation into methylation haplotype blocks could help identify mechanisms and contexts that give rise to the blocks. We also found significant overlap

of 5hmC with MHBs, therefore, it is possible that 5hmC signals might confound MHBs identification. oxBS-seq datasets, which provide true 5mC signals, can help resolve this issue. Furthermore, improving sample preparation and library construction on cell free DNA could greatly advance the development of a clinical assay using cell free DNA. First, due to the fragmented nature of cell free DNA, application of scRRBS protocol to cell free DNA only enriched for RRBS targets by two folds whereas on unfragmented DNA, the enrichment is typically by at least twenty folds. One possible solution is to perform post amplification capture of library fragments containing targeted regions.

In summary, this work have advanced DNA methylation biomarkers developments while demonstrating the successful integration of bioinformatics and biomolecular techniques. With further application of these tools to the development of more specific and robust clinical DNA biomarkers we can begin to diagnose, make prognosis, and track human diseases.

REFERENCES

1. Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Le Calvez-Kelm, F., Kaaks, R., Barrdahl, M., Boeing, H., Aleksandrova, K., Trichopoulou, A., Lagiou, P., Naska, A., Palli, D., Krogh, V., Polidoro, S., Tumino, R., Panico, S., Bueno-de-Mesquita, B., Peeters, P. H., Quirós, J. R., Navarro, C., Ardanaz, E., Dorronsoro, M., Key, T., Vineis, P., Murphy, N., Riboli, E., Romieu, I. & Herceg, Z. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* **8,** 599–618 (2016).

2. Lim, U. & Song, M.-A. Dietary and lifestyle factors of DNA methylation. *Methods Mol. Biol.* **863,** 359–376 (2012).

3. Mitchell, C., Schneper, L. M. & Notterman, D. A. DNA methylation, early life environment, and health outcomes. *Pediatr. Res.* **79,** 212–219 (2016).

4. Heard, E. & Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157,** 95–109 (2014).

5. Wigler, M., Levy, D. & Perucho, M. The somatic replication of DNA methylation. *Cell* **24,** 33–40 (1981).

6. Zhu, J.-K. Active DNA demethylation mediated by DNA glycosylases. *Annu. Rev. Genet.* **43,** 143–166 (2009).

7. Schübeler, D. Function and information content of DNA methylation. *Nature* **517,** 321–326 (2015).

8. Thienpont, B., Steinbacher, J., Zhao, H., D'Anna, F., Kuchnio, A., Ploumakis, A., Ghesquière, B., Van Dyck, L., Boeckx, B., Schoonjans, L., Hermans, E., Amant, F., Kristensen, V. N., Koh, K. P., Mazzone, M., Coleman, M. L., Carell, T., Carmeliet, P. & Lambrechts, D. Tumour hypoxia causes DNA hypermethylation by reducing TET activity. *Nature* **537,** 63–68 (2016).

9. Zhang, F., Liu, Y., Zhang, Z., Li, J., Wan, Y., Zhang, L., Wang, Y., Li, X., Xu, Y., Fu, X., Zhang, X., Zhang, M., Zhang, Z., Zhang, J., Yan, Q., Ye, J., Wang, Z., Chen, C. D., Lin, W. & Li, Q. 5-hydroxymethylcytosine loss is associated with poor prognosis for patients with WHO grade II diffuse astrocytomas. *Sci Rep* **6,** 20882 (2016).

10. Johnson, K. C., Houseman, E. A., King, J. E., von Herrmann, K. M., Fadul, C. E. & Christensen, B. C. 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nat Commun* **7,** 13177 (2016).

11. Chen, K., Zhang, J., Guo, Z., Ma, Q., Xu, Z., Zhou, Y., Xu, Z., Li, Z., Liu, Y., Ye, X., Li, X., Yuan, B., Ke, Y., He, C., Zhou, L., Liu, J. & Ci, W. Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Res.* **26,** 103–118 (2016).

12. He, Y. & Ecker, J. R. Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet* **16,** 55–77 (2015).

13. Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., Yen, C., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E., Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R. & Ren, B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153,** 1134–1148 (2013).

14. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.* **14,** 341–356 (2013).

15. He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C. & Xu, G.-L. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).

16. Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C. & Zhang, Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–1303 (2011).

17. Seisenberger, S., Peat, J. R. & Reik, W. Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. *Curr. Opin. Cell Biol.* **25,** 281–288 (2013).

18. Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A. C., Fung, H.-L., Zhang, K. & Zhang, Y. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153,** 692–706 (2013).

19. Song, C.-X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S.-Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., Gao, J., Liu, P., Li, L., Xu, G.-L., Jin, P. & He, C. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153,** 678–691 (2013).

20. Raiber, E.-A., Beraldi, D., Ficz, G., Burgess, H. E., Branco, M. R., Murat, P., Oxley, D., Booth, M. J., Reik, W. & Balasubramanian, S. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13,** R69 (2012).

21. Iurlaro, M., McInroy, G. R., Burgess, H. E., Dean, W., Raiber, E.-A., Bachman, M., Beraldi, D., Balasubramanian, S. & Reik, W. In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.* **17,** 141 (2016).

22. Jin, S.-G., Zhang, Z.-M., Dunwell, T. L., Harter, M. R., Wu, X., Johnson, J., Li, Z., Liu, J., Szabó, P. E., Lu, Q., Xu, G., Song, J. & Pfeifer, G. P. Tet3 Reads 5-

Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. *Cell Rep* **14,** 493–505 (2016).

23. Ngo, T. T. M., Yoo, J., Dai, Q., Zhang, Q., He, C., Aksimentiev, A. & Ha, T. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat Commun* **7,** 10813 (2016).

24. Hon, G. C., Song, C.-X., Du, T., Jin, F., Selvaraj, S., Lee, A. Y., Yen, C.-A., Ye, Z., Mao, S.-Q., Wang, B.-A., Kuan, S., Edsall, L. E., Zhao, B. S., Xu, G.-L., He, C. & Ren, B. 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol. Cell* **56,** 286–297 (2014).

25. Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. A., Marques, C. J., Andrews, S. & Reik, W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473,** 398–402 (2011).

26. Mikeska, T., Bock, C., Do, H. & Dobrovic, A. DNA methylation biomarkers in cancer: progress towards clinical implementation. *Expert Rev. Mol. Diagn.* **12,** 473–487 (2012).

27. Levenson, V. V. DNA methylation as a universal biomarker. *Expert Rev. Mol. Diagn.* **10,** 481–488 (2010).

28. Issa, J.-P. J. & Kantarjian, H. M. Targeting DNA methylation. *Clin. Cancer Res.* **15,** 3938–3946 (2009).

29. Wen, L., Li, J., Guo, H., Liu, X., Zheng, S., Zhang, D., Zhu, W., Qu, J., Guo, L., Du, D., Jin, X., Zhang, Y., Gao, Y., Shen, J., Ge, H., Tang, F., Huang, Y. & Peng, J. Genome-scale detection of hypermethylated CpG islands in circulating cell-free DNA of hepatocellular carcinoma patients. *Cell Res.* **25,** 1250–1264 (2015).

30. Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., Rubertsson, S., Nellgård, B., Blennow, K., Zetterberg, H., Spalding, K., Haller, M. J., Wasserfall, C. H., Schatz, D. A., Greenbaum, C. J., Dorrell, C., Grompe, M., Zick, A., Hubert, A., Maoz, M., Fendrich, V., Bartsch, D. K., Golan, T., Ben Sasson, S. A., Zamir, G., Razin, A., Cedar, H., Shapiro, A. M. J., Glaser, B., Shemer, R. & Dor, Y. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. U.S.A.* **113,** E1826-1834 (2016).

31. Sun, K., Jiang, P., Chan, K. C. A., Wong, J., Cheng, Y. K. Y., Liang, R. H. S., Chan, W., Ma, E. S. K., Chan, S. L., Cheng, S. H., Chan, R. W. Y., Tong, Y. K., Ng, S. S. M., Wong, R. S. M., Hui, D. S. C., Leung, T. N., Leung, T. Y., Lai, P. B. S., Chiu, R. W. K. & Lo, Y. M. D. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E5503-5512 (2015).

32. Huang, Y., Pastor, W. A., Shen, Y., Tahiliani, M., Liu, D. R. & Rao, A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5,** e8888 (2010).

33. Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48,** 226–232 (2009).

34. Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Khrebtukova, I., Milne, T. A., Huang, Y., Biswas, D., Hess, J. L., Allis, C. D., Roeder, R. G., Valk, P. J. M., Löwenberg, B., Delwel, R., Fernandez, H. F., Paietta, E., Tallman, M. S., Schroth, G. P., Mason, C. E., Melnick, A. & Figueroa, M. E. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet.* **8,** e1002781 (2012).

35. Plongthongkum, N., Diep, D. H. & Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.* **15,** 647–661 (2014).

36. Yu, M., Hon, G. C., Szulwach, K. E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J.-H., Jin, P., Ren, B. & He, C. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–1380 (2012).

37. Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W. & Balasubramanian, S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336,** 934–937 (2012).

38. Houseman, E. A., Johnson, K. C. & Christensen, B. C. OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. *Bioinformatics* **32,** 2505–2507 (2016).

39. Hahn, M. A., Li, A. X., Wu, X. & Pfeifer, G. P. Single base resolution analysis of 5-methylcytosine and 5-hydroxymethylcytosine by RRBS and TAB-RRBS. *Methods Mol. Biol.* **1238,** 273–287 (2015).

40. Nazor, K. L., Boland, M. J., Bibikova, M., Klotzle, B., Yu, M., Glenn-Pratola, V. L., Schell, J. P., Coleman, R. L., Cabral-da-Silva, M. C., Schmidt, U., Peterson, S. E., He, C., Loring, J. F. & Fan, J.-B. Application of a low cost array-based technique - TAB-Array - for quantifying and mapping both 5mC and 5hmC at single base resolution in human pluripotent stem cells. *Genomics* **104,** 358–367 (2014).

41. Field, S. F., Beraldi, D., Bachman, M., Stewart, S. K., Beck, S. & Balasubramanian, S. Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). *PLoS ONE* **10,** e0118202 (2015).

42. Ruiz, S., Diep, D., Gore, A., Panopoulos, A. D., Montserrat, N., Plongthongkum, N., Kumar, S., Fung, H.-L., Giorgetti, A., Bilic, J., Batchelder, E. M., Zaehres, H., Kan, N. G., Schöler, H. R., Mercola, M., Zhang, K. & Izpisua Belmonte, J. C. Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 16196–16201 (2012).

43. Yamaguchi, S., Hong, K., Liu, R., Shen, L., Inoue, A., Diep, D., Zhang, K. & Zhang, Y. Tet1 controls meiosis by regulating meiotic gene expression. *Nature* **492,** 443–447 (2012).

44. Park, Y. & Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32,** 1446–1453 (2016).

45. Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A. & Li, W. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15,** R38 (2014).

46. Dolzhenko, E. & Smith, A. D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **15,** 215 (2014).

47. Park, Y., Figueroa, M. E., Rozek, L. S. & Sartor, M. A. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* **30,** 2414–2422 (2014).

48. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. - PubMed - NCBI. at <https://www.ncbi.nlm.nih.gov/pubmed/23034086>

49. Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W. & Ecker, J. R. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–216 (2015).

50. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13,** R83 (2012).

51. Kishore, K., de Pretis, S., Lister, R., Morelli, M. J., Bianchi, V., Amati, B., Ecker, J. R. & Pelizzola, M. methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. *BMC Bioinformatics* **16,** 313 (2015).

52. Lee, W. & Morris, J. S. Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics* **32,** 664–672 (2016).

53. Wang, H., He, C., Kushwaha, G., Xu, D. & Qiu, J. A full Bayesian partition model for identifying hypo- and hyper-methylated loci from single nucleotide resolution sequencing data. *BMC Bioinformatics* **17 Suppl 1,** 7 (2016).

54. Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E. M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I.-H., Yu, J., Daley, G. Q., Eggan, K., Hochedlinger, K., Thomson, J., Wang, W., Gao, Y. & Zhang, K. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27,** 353–360 (2009).

55. Liu, G.-H., Barkho, B. Z., Ruiz, S., Diep, D., Qu, J., Yang, S.-L., Panopoulos, A. D., Suzuki, K., Kurian, L., Walsh, C., Thompson, J., Boue, S., Fung, H. L., Sancho-Martinez, I., Zhang, K., Yates, J. & Izpisua Belmonte, J. C. Recapitulation of premature ageing with iPSCs from Hutchinson-Gilford progeria syndrome. *Nature* **472,** 221–225 (2011).

56. Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A. J., Zheng, L., Zhang, H., Huang, S., Min, J., Nicholson, T., Chen, T., Xu, G., Shi, Y., Zhang, K. & Shi, Y. G. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol. Cell* **42,** 451–464 (2011).

57. Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A. & Feinberg, A. P. Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43,** 768–775 (2011).

58. Plongthongkum, N., van Eijk, K. R., de Jong, S., Wang, T., Sul, J. H., Boks, M. P. M., Kahn, R. S., Fung, H.-L., Ophoff, R. A. & Zhang, K. Characterization of genome-methylome interactions in 22 nuclear pedigrees. *PLoS ONE* **9,** e99313 (2014).

59. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20,** 883–889 (2010).

60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

61. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

62. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6,** 315–316 (2009).

63. Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S. & Feinberg, A. P. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41,** 178–186 (2009).

64. Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., Miller, J., Schlaeger, T., Daley, G. Q. & Feinberg, A. P. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41,** 1350–1353 (2009).

65. 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., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. & Ecker, J. R. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).

66. Figueroa, M. E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P. J., Schifano, E., Booth, J., van Putten, W., Skrabanek, L., Campagne, F., Mazumdar, M., Greally, J. M., Valk, P. J. M., Löwenberg, B., Delwel, R. & Melnick, A.

DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17,** 13–27 (2010).

67. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).

68. Huang, K., Shen, Y., Xue, Z., Bibikova, M., April, C., Liu, Z., Cheng, L., Nagy, A., Pellegrini, M., Fan, J.-B. & Fan, G. A Panel of CpG Methylation Sites Distinguishes Human Embryonic Stem Cells and Induced Pluripotent Stem Cells. *Stem Cell Reports* **2,** 36–43 (2013).

69. Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., Ziller, M., Croft, G. F., Amoroso, M. W., Oakley, D. H., Gnirke, A., Eggan, K. & Meissner, A. Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines. *Cell* **144,** 439–452 (2011).

70. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., FANTOM Consortium, Forrest, A. R. R., Carninci, P., Rehli, M. & Sandelin, A. An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

71. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).

72. Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. D. & Ren, B. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.* **45,** 1198–1206 (2013).

73. Liu, H., Liu, X., Zhang, S., Lv, J., Li, S., Shang, S., Jia, S., Wei, Y., Wang, F., Su, J., Wu, Q. & Zhang, Y. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res.* **44,** 75–94 (2016).

74. Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A. & Meissner, A. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).

75. Libertini, E., Heath, S. C., Hamoudi, R. A., Gut, M., Ziller, M. J., Czyz, A., Ruotti, V., Stunnenberg, H. G., Frontini, M., Ouwehand, W. H., Meissner, A., Gut, I. G. & Beck, S. Information recovery from low coverage whole-genome bisulfite sequencing. *Nature Communications* **7,** 11306 (2016).

76. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–663 (2007).

77. Young-Xu, Y. & Chan, K. A. Pooling overdispersed binomial data to estimate event rate. *BMC Med Res Methodol* **8,** 58 (2008).

78. Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., Kotliar, D., Zhang, W., Ghandi, M., Garraway, L., Fernandes, S. M., Livak, K. J., Gabriel, S., Gnirke, A., Lander, E. S., Brown, J. R., Neuberg, D., Kharchenko, P. V., Hacohen, N., Getz, G., Meissner, A. & Wu, C. J. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26,** 813–825 (2014).

79. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9,** 477–485 (2008).

80. Li, S., Garrett-Bakelman, F., Perl, A. E., Luger, S. M., Zhang, C., To, B. L., Lewis, I. D., Brown, A. L., D'Andrea, R. J., Ross, M. E., Levine, R., Carroll, M., Melnick, A. & Mason, C. E. Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biol.* **15,** 472 (2014).

81. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nat. Rev. Genet.* **16,** 716–726 (2015).

82. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167,** 1145–1149 (2016).

83. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

84. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S. & Thomson, J. A. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–1048 (2010).

85. Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T. & Marsit, C. J. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17,** 259 (2016).

86. Saito, D. & Suyama, M. Linkage disequilibrium analysis of allelic heterogeneity in DNA methylation. *Epigenetics* **10,** 1093–1098 (2015).

87. Heyn, H., Li, N., Ferreira, H. J., Moran, S., Pisano, D. G., Gomez, A., Diez, J., Sanchez-Mut, J. V., Setien, F., Carmona, F. J., Puca, A. A., Sayols, S., Pujana, M. A., Serra-Musach, J., Iglesias-Platas, I., Formiga, F., Fernandez, A. F., Fraga, M. F., Heath, S. C., Valencia, A., Gut, I. G., Wang, J. & Esteller, M. Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 10522–10527 (2012).

88. Blattler, A., Yao, L., Witt, H., Guo, Y., Nicolet, C. M., Berman, B. P. & Farnham, P. J. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol.* **15,** 469 (2014).

89. Heyn, H., Vidal, E., Ferreira, H. J., Vizoso, M., Sayols, S., Gomez, A., Moran, S., Boque-Sastre, R., Guil, S., Martinez-Cardus, A., Lin, C. Y., Royo, R., Sanchez-Mut, J. V., Martinez, R., Gut, M., Torrents, D., Orozco, M., Gut, I., Young, R. A. & Esteller, M. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* **17,** 11 (2016).

90. Shao, X., Zhang, C., Sun, M.-A., Lu, X. & Xie, H. Deciphering the heterogeneity in DNA methylation patterns during stem cell differentiation and reprogramming. *BMC Genomics* **15,** 978 (2014).

91. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W. & van Steensel, B. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

92. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat. Genet.* **41,** 246–250 (2009).

93. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

94. Pujadas, E. & Feinberg, A. P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148,** 1123–1131 (2012).

95. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L., Yang, H., Chi, N. C., Zhang, M. Q., Ecker, J. R. & Ren, B. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350–354 (2015).

96. Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M. & Yamanaka, S. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113,** 631–642 (2003).

97. Shu, J., Wu, C., Wu, Y., Li, Z., Shao, S., Zhao, W., Tang, X., Yang, H., Shen, L., Zuo, X., Yang, W., Shi, Y., Chi, X., Zhang, H., Gao, G., Shu, Y., Yuan, K., He, W., Tang, C., Zhao, Y. & Deng, H. Induction of pluripotency in mouse somatic cells with lineage specifiers. *Cell* **153,** 963–975 (2013).

98. 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.* **23,** 2126–2135 (2013).

99. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164,** 57–68 (2016).

100. Williams, K., Christensen, J., Pedersen, M. T., Johansen, J. V., Cloos, P. A. C., Rappsilber, J. & Helin, K. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473,** 343–348 (2011).

101. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 3740–3745 (2002).

102. Timmons, J. A., Szkop, K. J. & Gallagher, I. J. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol.* **16,** 186 (2015).

103. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A. & Young, R. A. Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).

104. Xie, H., Wang, M., de Andrade, A., Bonaldo, M. de F., Galat, V., Arndt, K., Rajaram, V., Goldman, S., Tomita, T. & Soares, M. B. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.* **39,** 4099–4108 (2011).

105. Landan, G., Cohen, N. M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D. A., Goldfinger, N., Zundelevich, A., Gal-Yam, E. N., Rotter, V. & Tanay, A. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44,** 1207–1214 (2012).

106. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

107. Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. & Kelsey, K. T. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).

108. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29,** 1083–1085 (2013).

109. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* **13,** 497–510 (2013).

110. Sproul, D., Kitchen, R. R., Nestor, C. E., Dixon, J. M., Sims, A. H., Harrison, D. J., Ramsahoye, B. H. & Meehan, R. R. Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol.* **13,** R84 (2012).

111. Uribe-Lewis, S., Stark, R., Carroll, T., Dunning, M. J., Bachman, M., Ito, Y., Stojic, L., Halim, S., Vowler, S. L., Lynch, A. G., Delatte, B., de Bony, E. J., Colin, L., Defrance, M., Krueger, F., Silva, A.-L., Ten Hoopen, R., Ibrahim, A. E., Fuks, F. & Murrell, A. 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. *Genome Biol.* **16,** 69 (2015).

112. Deplus, R., Delatte, B., Schwinn, M. K., Defrance, M., Méndez, J., Murphy, N., Dawson, M. A., Volkmar, M., Putmans, P., Calonne, E., Shih, A. H., Levine, R. L., Bernard, O., Mercher, T., Solary, E., Urh, M., Daniels, D. L. & Fuks, F. TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *EMBO J.* **32,** 645–655 (2013).

113. Hohos, N. M., Lee, K., Ji, L., Yu, M., Kandasamy, M. M., Phillips, B. G., Baile, C. A., He, C., Schmitz, R. J. & Meagher, R. B. DNA cytosine hydroxymethylation levels are distinct among non-overlapping classes of peripheral blood leukocytes. *J. Immunol. Methods* **436,** 1–15 (2016).

114. Pacis, A., Tailleux, L., Morin, A. M., Lambourne, J., MacIsaac, J. L., Yotova, V., Dumaine, A., Danckaert, A., Luca, F., Grenier, J.-C., Hansen, K. D., Gicquel, B., Yu, M., Pai, A., He, C., Tung, J., Pastinen, T., Kobor, M. S., Pique-Regi, R., Gilad, Y. & Barreiro, L. B. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res.* **25,** 1801–1811 (2015).

115. Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C., Yu, M., Liu, X., Zhu, P., Li, X., Hou, Y., Guo, H., Wu, X., He, C., Li, R., Tang, F. & Qiao, J. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol.* **15,** R49 (2014).

116. Jensen, T. J., Kim, S. K., Zhu, Z., Chin, C., Gebhard, C., Lu, T., Deciu, C., van den Boom, D. & Ehrich, M. Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. *Genome Biol.* **16,** 78 (2015).

APPENDIX

Table A-1. List of BSPP primers

Primer name	Primer sequences
pAP1V61U	5’-G*G*G*TCATATCGGTCACTGTU-3’
AP2V6	5’-/5Phos/CACGGGTAGTGTGTATCCTG-3’
RE-DpnII-V6	5’-GTGTATCCTGATC-3’
AmpF6.4Sol	5’-AATGATACGGCGACCACCGAGATCTACACCACTCTCAGATGTTATCGAGGTCCGAC-3’
AmpF6.3NH2	5’-/5AmMC6/CAGATGTTATCGAGGTCCGAC-3’
AmpR6.3NH2	5’-/5AmMC6/GGAACGATGAGCCTCCAAC-3’
PCR_F	5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG CTCTTC-3’
PE_t_N2	5’-ACACTCTTTCCCT ACACGACGCTCTTCCGA TCTN*N-3’
PE_b_A	5’-/5Phos/AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3’
eMIP_CA1_F	5’- TGCCTAGGACCGGATCAACT-3’
eMIP_CA1_R	5’- GAGCTTCGGTTCACGCAATG-3’
CP-2-FA	5’-GCACGATCCGACGGTAGTGT-3’
CP-2-RA	5’-CCGTAATCGGGAAGCTGAAG-3’
CA-2-FA.Indx7Sol	5’-CAAGCAGAAGACGGCATACGAGATGATCTGCGGTCTGCCATCCGACGGTAGTGT-3’
CA-2-FA.Indx45Sol	5’-CAAGCAGAAGACGGCATACGAGATCGTAGTCGGTCTGCCATCCGACGGTAGTGT-3’
CA-2-FA.Indx76Sol	5’-CAAGCAGAAGACGGCATACGAGATAATAGGCGGTCTGCCATCCGACGGTAGTGT-3’
CA-2-RA.Sol	5’- AATGATACGGCGACCACCGAGATCTACACGCCTATCGGGAAGCTGAAG-3’
Switch.CA-2-FA.Sol	5’- AATGATACGGCGACCACCGAGATCTACACGCCTATCCGACGGTAGTGT-3’
Switch.CA-2-RA.Ind7Sol	5’-CAAGCAGAAGACGGCATACGAGATGATCTGCGGTCTGCCATCGGGAAGCTGAAG-3’
Switch.CA-2-RA.Ind45Sol	5’-CAAGCAGAAGACGGCATACGAGATCGTAGTCGGTCTGCCATCGGGAAGCTGAAG-3’
Switch.CA-2-RA.Ind76Sol	5’-CAAGCAGAAGACGGCATACGAGATAATAGGCGGTCTGCCATCGGGAAGCTGAAG-3’

* Indicates a phosphorothioate bond

Table A-2. Lambda DNA primers for generating control DNA

Primer ID	Primer seq	Purpose	Direction	Amplicon size
Lambda_hmC_F	CAGGAAGACAGTGCTCATGC	5hmC	F	189
Lambda_hmC_R	CCAGCAGGGATTTCTCCTGT	5hmC	R	
bisLambda_hmC_F	TAGGAAGATAGTGTTTATGT	5hmC	F	189
bisLambda_hmC_R	CCAACAAAAATTTCTCCTAT	5hmC	R	
Lambda_mC_F	TGTTATTCATGTTGCATGGTGC	5mC	F	262
Lambda_mC_R	CAGCTGACTTCTTTTCTTTCAC	5mC	R	
bisLambda_mC_F	TGTTATTTATGTTGTATGGTGT	5mC	F	262
bisLambda_mC_R	CAACTAACTTCTTTTCTTTCAC	5mC	R	
Lambda_C_F	TTGCTCATAGGAGATATGGT	C	F	277
Lambda_C_R	CTTGCTAACCAATTCCTAGG	C	R	
bisLambda_C_F	TTGTTTATAGGAGATATGGTAGA	C	F	277
bisLambda_C_R	CTTACTAACCAATTCCTAAA	C	R	

Designing bisulfite padlock probes with ppDesigner

Reference files and dependencies

1. *ppDesigner* can be downloaded from http://genome-tech.ucsd.edu/public/Gen2_BSPP/ppDesigner/ppDesigner.php.
2. A Mac OS or any other modern Unix-based system is required.
3. *Perl* is required. It should already be included in all Unix-based system.
4. *Perl* modules, File::Temp and Sort::Array, are required to run *ppDesigner* and can be downloaded from http://www.cpan.org.
5. *BioPerl* toolkit is required to run *ppDesigner* and can be obtained from http://www.bioperl.org.
6. **Optional:** *UNAfold* software version 3.8. Using *UNAFold* will result in a more accurate prediction of probe efficiency. It is not required, and it will not change the probe sequence.
7. Genome reference sequences in FASTA format is required and can be downloaded from UCSC Genome Browser. These cannot be in the multi-FASTA format with multiple chromosomes per file.
8. Targeted region list in BED format is required to design targets.

Step-by-step

1. Unzip the *ppDesigner* software package.
2. Ensure that individual reference sequences (FASTA) files are placed in a common directory.
3. Convert the target list BED file to a target file in the format required. The file should be tab-delimited and have four required columns (1) the unique ID for each target region, (2) the FASTA filename, such as chr22 (for chr22.fa), (3) The starting position, and (4) The ending position. The final fifth column can be the required strand to capture. If strand is not indicated, the program will pick the more efficient probes from either strand. An example target file is given in the Example folder.
4. Generate a job file in the format required. An example job file is given in the Example folder. All of the parameters must be given. The **unafold_path** variable can be set to 'NA' if **using_unafold=0**. All paths must be full paths. See the **Notes** section for important considerations in choosing parameter values.
5. Run ppDesigner.pl script as follows. See the README for *ppDesigner* for specific usage instructions.

```
/ppDesigner_BSPP_v2.0/src/ppDesigner.pl jobFile.pl >
Outputs.txt
```

6. Run the primer2padlock.pl script as follows. The **maxH1H2len** is a numerical value and should be the same as the **H1_plus_H2_len** variable from the job file. Indicate "**array**" to generate probes sequences that contain amplification adapters for array synthesis of probes.

```
/ppDesigner_BSPP_v2.0/src/primer2padlock.pl maxH1H2len
[array or empty] < Outputs.txt > probes.seq
```

7. The probe sequences are now ready to be synthesized. It is recommended to randomize the order of the sequences so that technical effects such as bad quality spots on the arrays will affect probes in a random manner do not appear as systematic errors.

Notes

- The maximum target length must take into consideration the desired sequencing platform for the assay. For most applications, a maximum target length of 100bp should be sufficient. Longer target lengths will require longer sequencing reads or paired-end sequencing to cover the entire targeted region. There is also an inverse correlation between the capture efficiency and the target length.
- The minimum and maximum target length must be close to enable selection of specific capture products prior to sequencing. Larger differences will lead to more potential non-specific products being sequenced.
- The maximum number of CpGs in the capture arm should be limited to 0 or 1. More CpGs means that more alternative probes must be synthesized to avoid capture bias towards methylated or unmethylated targets. The capture efficiency of probes can also vary between methylated or unmethylated probes. The presence of CH methylation can be negligible in most cell types so we safely assume that they are unmethylated.
- Minimum H1 and H2 lengths lower than 25 should be avoided as this may lead to capture sequences with very low melting temperatures that may not be able to anneal efficiently. H1 plus H2 length should not be longer than 60 as this will lead to higher synthesis cost and potentially more non-specific products.
- Unique molecular identifiers (UMI) can be attached to padlock probes to allow for single molecules counting and removal of clonal read

Performing bisulfite reads analysis with BisReadMapper

Reference files and dependencies

1. BisReadMapper can be downloaded from github: *https://github.com/dinhdiep/BisReadMapper*
2. A Mac OS or any other modern Unix-based system is required.
3. *Perl* is required. It should already be included in all Unix-based system.
4. Trim Galore! is required: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ Note that Cutadapt is required also for Trim Galore!
5. bamUtils is required: http://genome.sph.umich.edu/wiki/BamUtil
6. One of the four supported aligners: Supported aligners are bowtie2 (version bowtie2-2.1.0), bwa (bwa-0.7.5a), SOAP2 (soap2.21release) , LAST (last-458) , or GEM (GEM-binaries-Linux-x86_64-core_i3-20130406-045632). Note that BWA is recommended for general usage.
7. Perl module Statistics::LSNoHistory is required for computing the Pearson's correlation between Watson and Crick strands.
8. Samtools is required: http://samtools.sourceforge.net/
9. Genome reference files in FASTA format
10. Genome reference chromosome sizes file.

Step-by-step

1. Use genomePrep.pl to generate the in-silico bisulfite converted references, C->T for Watson, and G->A for Crick strands. Note that bisulfite conversion makes the two strands non-complementary.
2. Use aligner software to create index files from the reference sequence for alignment. NOTE: Both strands (*.bis.CT and *.bis.GA) can be concatenated into one file and only one index needs to be created so long as the aligner can support larger index files.
3. Use samtools to generate the *.fai file from the reference sequence.
4. Generate list_fastq_file or a table of the files to be processed. Each column in this file represents:
   <sample id> <dir> <read1.fq | read1.fq,read2.fq | *.sam> <phred> <clonal method> <adaptor r1 sequence> <adaptor r2 sequence>
5. Generate the list_paths file
6. Run MasterBisReadMapper.pl with list_fastq_files and list_paths as inputs.

Identifying differential methylation with cgDMR-miner

Reference files and dependencies

1. *cgDMR-miner* can be downloaded from *https://github.com/dinhdiep/cgDMR-miner*
2. A Mac OS or any other modern Unix-based system is required.
3. *Perl* is required. It should already be included in all Unix-based system.
4. *R* is required. It can be downloaded from https://www.r-project.org/
5. *Perl* modules, Math::Random, Statistics::LSNoHistory, Statistics::Basic, Statistics::Descriptive, are required and can be downloaded from http://www.cpan.org.
6. The following R packages are required and can be downloaded from Bioconductor: bsseq, data.table, depmixS4, DNAcopy, entropy, fastseg, GenomicRanges, graphics, MASS, methods, mgcv, mhsmm, pryr, RColorBrewer.
7. Reference genome CpG position lists. The genomePrep.pl script from BisReadMapper can be used to obtain CpG positions from any reference FASTA file, then converted to BED using the following awk command.

```
sed 's/:/\t/g' cpg.positions.txt | awk '{printf("%s\t%d\t%d\n",
$1, $3-1, $3)}' > cpg.positions.bed
```

Usage

```
perl cgDMR-miner.pl samplesInfo cpgBedList minDepth pValueCutoff
minEffectSize segmentationMode
```

samplesInfo. A tab separated file with three columns and one row per sample, (1) sample name, (2) sample id, (3) path to list of methylation frequency files. For each sample a list of methylation frequency file must be generated, which is a tab separated table comprised of three columns and one row per chromosome: (1) sample name matching samples_info file, (2) path to the specific

methylation frequency file, (3) chromosome name. Note that both sample names and sample ids should not contain any spaces. See Example.

cpgBedList. A tab separated file with two columns, (1) chromosome name, (2) path to cpg positions bed file. For each chromosome, a bed file containing the CpG positions to be considered must be provided. If chromosome bed file is missing, then that chromosome will be ignored. Note that chromosome names should not contain any spaces. See Example.

minDepth. An integer that is the minimum total depth required in each sample for DMR summarization. Default is 10.

pValueCutoff. A floating value [0-1] that is the maximum p-value for DMRs. Default is 0.001

minEffectSize. a floating value greater or equal to 0 that is the effect size cutoff for DMR calling. This is equals to the square root of the test statistic divided by N (total depth of coverage for all samples).

segmentationMode. "HMM" or "CBS". HMM option will utilize 5-states Hidden Markov Model while CBS option will utilize circular binary segmentation. I recommend HMM for WGBS data and CBS for other genome-wide datasets such as RRBS or capture sequencing data.