Title

Probing Interaction of Genome and Methylome by Targeted Bisulfite Sequencing

Author

Plongthongkum, Nongluk

Publication Date
2014

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Probing Interaction of Genome and Methylome by Targeted Bisulfite Sequencing

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

Doctor of Philosophy

in

Bioengineering

by

Nongluk Plongthongkum

Committee in charge:

Professor Kun Zhang, Chair
Professor Kelly Frazer
Professor Jeff Hasty
Professor Xiaohua Huang
Professor Christopher Woelk

2014

The Dissertation of Nongluk Plongthongkum is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2014

DEDICATION

I dedicate my dissertation to my dearest parents, San Plongthongkum and Thongyoi Plongthongkum, for their unconditional love and support throughout my life. I also dedicate this dissertation to my siblings for their love and support in everyway.

EPIGRAPH

“Learn from yesterday, live for today, hope for tomorrow. The important thing is to not stop questioning” – Albert Einstein

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BSPP	Bisulfite padlock probe
MeDIP	Methylated DNA immunoprecipitation
MBD	Methyl-CpG-binding domain
PCR	Polymerase chain reaction
DNMT	DNA methylatransferase
TET	Ten-eleven translocation
C	Cytosine
T	Thymine
5mC	5-methylcytosine
5hmC	5-hydroxymethylcytosine
5fC	5-formylcytosine
5caC	5-carboxylcytosine
BS-seq	Bisulfite sequencing
WGBS	Whole-genome bisulfite sequencing
PBAT	Post-bisulfite adaptor tagging
RRBS	Reduced representation bisulfite sequencing
LHC-BS	Liquid hybridization capture-based bisulfite sequencing
RSMA	Restriction enzyme-based single-cell methylation assay
CHARM	Comprehensive high-throughput arrays for relative methylation
ChIP	Chromatin immunoprecipitation
MRE-seq	Methylation-sensitive restriction enzyme sequencing
SNP	Single-nucleotide polymorphism
PAGE	Polyacrylamide gel electrophoresis
CTCF	CCCTC-binding factor
ESC	Embryonic stem cell
iPSC	Induced pluripotent stem cell
DMS	Differentially methylated site
DMR	Differentially methylated region

FFPE	Formalin fixed paraffin embedded
UMI	Unique molecular identifier
mQTL	Methylation quantitative trait loci
ASM	Allele-specific methylation
MPO	Mid-parent offspring
VMR	Variably methylated region
vSNP	Variation-single-nucleotide polymorphism
FDR	False discovery rate
TSS	Transcription start site
UTR	Untranslated region
BH	Benjamini-Hochberg
STD	Standard deviation

LIST OF SUPPLEMENTARY FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude and appreciation to Dr. Kun Zhang for giving me an opportunity to join his lab and being a superb graduate advisor. I thank him for his guidance, encouragement, patience, and support while I was in his lab. Working with him is my greatest pleasure. He is undoubtedly the best mentor one could ask for. He always has a great vision and gave me a lot of opportunities to work in different projects, which is the best training I ever had as a Ph.D. student.

I would also like to thank all the members of the Zhang's labs. They are very supportive and being good companions. I would like to give my special thanks to Dinh Diep who is my first mentor when I joined the Zhang's lab. She always helps me a lot about computational stuffs, and she is a very good co-worker.

I would also like to thank my committee members, Dr. Jeff Hasty, Dr. Xiaohua Huang, Dr. Kelly Frazer, Dr. Christopher Woelk, for their precious time to serve as my committee members.

One of my friends that I cannot forget to acknowledge is Amy Chan. I thank her for being by my side while I was being in the tough time in the third year of my study. She gave me supports to get through the hardest time.

I would like to wish my deepest thanks to my ex-advisor back in Thailand, Dr. Witoon Tirasophon. He is one of the greatest advisors I ever had. He is the role model that inspired me to follow his footsteps to be a good scientist. Another person I would like to thank is Dr. Sakol Panyim. Without his support and encouragement, getting accepted into UCSD would not have been possible.

Financially, I would like to acknowledge the Office of the Higher Educational Commission, Royal Thai Government for the financial support of my Ph.D. from September 2007 to September 2012. In addition, I also would like to acknowledge my advisor, Dr. Kun Zhang, for financial support from 2011 to 2014.

Last but not least, I would like to thank the Bioengineering Department at UCSD, their faculty, and administration staffs. I would like to give my special thank to Jan Lenington, my graduate coordinator, for her helps and management of my documents for Ph.D. study.

Chapter 1, in part, is a reprint of the material as it appears Nature Review Genetics 2014 Vol. 15. Nongluk Plongthongkum, Dinh H. Diep, and Kun Zhang. "Advances in the profiling of DNA modifications: cytosine methylation and beyond." Nature Review Genetics 15, (2014): 647-661. The dissertation author was co-author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Nature Methods 2012 Vol. 9. Dinh Diep, Nongluk Plongthongkum, Athurva Gore, Ho-Lim Fung, Robert Shoemaker and Kun Zhang. "Library-free methylation sequencing with bisulfite padlock probe." Nature Methods 9(3), (2012): 270-272. The dissertation author was the primary investigator and co-author of this paper.

Chapter 3, in full, is the study that was performed as a part of BLUEPRINT Biomarker Benchmark project. The current status of this project is in the progress of analyzing additional data to assess the power of each assay and to validate the workflow and cost structure. The consortium paper will be released after all validations are finished. I acknowledge Dinh H. Diep for contribution in this work in part of padlock probe design and data processing. I was a primary author and performed BSPP assay.

Chapter 4, in full, is a reprint of the material as it appears in PLoS One 2014. Vol9. Nongluk Plongthongkum, Kristel R. van Eijk, Simone de Jong, Tina Wang, Jae Hoon Sul, Marco P.M. Boks, Rene S. Kahn, Ho-Lim Fung, Roel A. Ophoff, and Kun Zhang. Characterization of Genome-Methylome Interactions in 22 Nuclear Pedigrees. PLoS One 9(7), (2014):e99313. The dissertation author was the primary investigator and author of this paper.

VITA

2001	Bachelor of Sciences in Biotechnology King Mongkut's Instiitute of Technology Ladkrabang, Thailand
2001-2002	Research assistant Mahidol University, Thailand
2005	Master of Sciences in Molecular Genetic and Genetic Engineering Mahidol University, Thailand
2005-2007	Research assistant Mahidol University, Thailand
2014	Doctor of Philosophy in Bioengineering University of California, San Diego

List of Publications

1. **Plongthongkum N**, van Eijk KR, de Jong S, Wang T, Sul JH, Boks MP, Kahn RS, Fung HL, Ophoff RA, Zhang K. Characterization of genome-methylome interactions in 22 nuclear pedigrees. PLoS One 2014. 9(7):e99313.
2. **Plongthongkum* N,** Diep D*, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet 2014. 15(10):647-61. (*co-first author)
3. Ruiz S, Diep D, Gore A, Panopoulos AD, Montserrat N, **Plongthongkum N**, Kumar S, Fung HL, Giorgetti A, Bilic J, Batchelder EM, Zaehres H, Kan NG, Scholer HR, Mercola M, Zhang K, Izpisua Belmonte JC. Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. Proc Natl Acad Sci U S A 2012. 109(40):16196-201.
4. Diep D*, **Plongthongkum N***, Gore A*, Fung HL, Shoemaker R, Zhang K. Library-free methylation sequencing with bisulfite padlock probes. Nat Methods 2012. 9(3):270-2. (*co-first author)

ABSTRACT OF THE DISSERTATION

Probing Interaction of Genome and Methylome by Targeted Bisulfite Sequencing

by

Nongluk Plongthongkum

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2014

Professor Kun Zhang, Chair

DNA methylation at CpG dinucleotides in mammalian cells is recognized as an epigenetic mechanism that plays a major role in mammalian development via gene expression regulation. Techniques in DNA methylation profiling have been advancing in the past decades. I have developed the second-generation of bisulfite padlock probe (BSPP) method, which does not require multiple steps of standard library preparation. This method is high-throughput and more scalable for quantification of DNA methylation at single-base resolution. The library-free method greatly reduces sample-preparation time and cost and is also compatible with automation. These developments have fulfilled the key requirements of a DNA methylation assay, including cost effectiveness, minimum sample input requirements, accuracy, and throughput. I have performed this technique to compare with other assays performed by different research groups for locus-specific DNA methylation analysis on the same samples set. BSPP assay showed a high correlation with other assays that have highest accuracy and is at the top with other assays based on the throughput. Genetic variants have an impact on local DNA methylation patterns by influencing methyltransferase recognition sequences or altering the DNA binding affinity of *cis*-regulatory proteins. To study this interaction, I have characterized CpG methylation state of 96 individuals from 22 nuclear pedigrees consisting of 52 parent-child trios using BSPP. I used the DMR330k probe set to quantify DNA methylation level at a set of 411,800 CpGs. Next, I have employed three independent approaches, including mid-parent offspring (MPO), methylation quantitative trait loci (mQTL), and allele-specific methylation (ASM) analysis, to investigate the influence of genetic polymorphisms on DNA methylation variation. MPO analysis identified 10,593 heritable CpG sites, among which 70.1% were SNPs that present in CpG sites. With mQTL analysis, 49.9% of heritable CpG sites were identified where regulation occurred in a distal *cis*-regulatory manner while ASM analysis was only able to identify 5% of heritable CpGs. This finding suggested that mQTL analysis do not identify all the *cis*-regulartory SNPs associated with heritable CpG methylation, and ASM analysis has even less power. I have extensively proved that in addition to regulating the mean of DNA methylation, genetic polymorphisms are also associated with the variability of DNA methylation levels. I have identified hundreds of CpG

clusters in human genome for which the degree of DNA methylation variability was associated with genetic polymorphisms. This finding supported the previous studies showing that genetic variants have the influence on phenotypic plasticity such as gene expression or DNA methylation.

CHAPTER 1 INTRODUCTION

1.1 Basic of DNA methylation

DNA methylation, most commonly recognized as 5-methylcytosine (5mC), is a key epigenetic mark that has essential roles in cellular processes including gene transcriptional regulation, genomic imprinting, embryonic development, X-chromosome inactivation, and disease susceptibility or development. DNA methylation can be created and erased dynamically but can also be stably maintained through cell divisions. Whole genome maps of 5mC have revealed intriguing patterns in human and mouse such as cell state dependent occurrences of 5mC in contexts other than canonical CpGs and in partially-methylated domains (PMDs), and conserved regions depleted of 5mCs across mouse and human species. The commonly known DNA methyl-transferases (DNMTs), which are well known for depositing methyl-groups on cytosine to yield 5mC in CpG contexts (Figure 1.1), have been shown to deposit methyl-groups at non-CpG sites [1]. Generation and maintenance of non-CpG methylation appears to be tightly regulated, for such modifications are enriched in specific cell types, such as pluripotent cells and neural progenitors as well as in adolescent and adult cortex tissues [2-6]. By contrast, partially-methylated domains (PMDs) have been found predominantly in non-pluripotent cells and non-cortex tissue types [2, 4, 7]. These PMDs have been associated with low transcription rates, lamina-associated domains and late-replicating domains. Next, different classes of methylation depleted regions named unmethylated regions (UMRs), DNA methylation valleys (DMVs) and DNA methylation canyons have been defined [7-9]. These regions tend to be conserved across cell types and across mouse and human species. Both methylation valleys and canyons tend to be marked with H3K4me3 or

H3K27me3 or both that can each lead to active, inactive, or poised transcriptional states respectively[7, 9, 10]. Strikingly, these regions cover most genes important for embryonic development [10].

In addition to DNMTs, a class of enzymes has been recently described to produce epigenetic modifications such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in mammalian cells. These newly identified writers are members of the ten-eleven translocation (TET) proteins and can sequentially oxidize 5mC to form 5hmC, 5fC and 5caC, respectively (Figure 1.1). These cytosine modifications will be referred to as 5mC oxidation derivatives from here on. Some reviews have suggested that 5mC oxidation derivatives may exist as demethylation intermediates and that their presence may be related to the development and maintenance of methylation-free regulatory regions in mammalian genomes [11, 12]. The presence of 5mC at major satellites and other transposable elements (TEs) has been reported to be necessary for genome stability [11] while depletion of 5mC in a small number of TEs are tissue-specific and may lead to enhancer functions [13]. 5hmC has been detected at short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs) [14-17], whereas 5fC and 5caC have been identified at major satellites [16]. These findings point to potential 5mC turnover at these regions. However, further investigation is required to determine the exact function of these DNA modifications. Proteins that preferentially bind to 5hmC, 5fC, or 5caC have been recently identified [18, 19] and are suggested to be the readers that connect these rare DNA modifications to phenotypic consequences. These exciting discoveries on DNA modifications were only possible with the advancement of techniques for characterizing these modifications as well as with the development of computational approaches for interpreting increasingly large datasets.

1.2 Measurement of DNA methylation

Techniques in DNA methylation measurement have been tremendously developed. A variety of methods has been advanced and optimized to measure DNA methylation at genome-wide scale or at specific regions of the genome. The four key requirements for 5mC measurement are improving accuracy, reducing sample input, increasing throughput, and lowering cost. The advancement of DNA methylation detection by chemical treatment with sodium bisulfite and sequencing-based method (bisulfite sequencing, BS-seq) has been very active because next-generation sequencing is becoming more affordable and can provide quantification in the form of digital counts, enabling merging of data from different batches or sequencing run of sequencing libraries, as well as for data from independent studies. With ordinary DNA sequencing method, unmethylated cytosine and methylated cytosine can not be distinguished. In addition, DNA methylation signal is erased after DNA amplification. Therefore DNA treatment by sodium bisulfite is used to convert unmethylated cytosine into uracil by sulfonation, deamination, and desulfonation procedures (Figure 1.2). After polymerase chain reaction, unmethylated cytosine is readout as thymine while methylated cytosine that resists conversion is unchanged. Quantification of 5mC at individual position is the ratio of methylated cytosines over total cytosines called. The majority of sequencing-based methods could be subcategorized as whole-genome methods, non-targeted enrichment methods, and targeted enrichment methods (Table 1.1). The whole-genome methods provide the pattern of DNA methylation across the whole genome except the methylation at repetitive regions, as the reads originating from repeats are assigned to multiple genomic regions and discarded. Non-targeted enrichment and targeted enrichment methods provide DNA methylation at specific subset of CpG sites in the genome based on custom design or the enrichment approaches. The low-input improvement to as low as 100pg or in single cell level allows the assays applicable for detecting DNA methylation in rare cell types, such as primordial germ cells and oocytes.

Whole-genome bisulfite sequencing (WGBS) has been considered as a ‘gold standard method’ in DNA methylation profiling, as it can profile DNA methylation at every single cytosine at single-base resolution across the entire genome. The basic procedures of WGBS method are fragmentation of gDNA followed by addition of adaptor, bisulfite treatment, and amplification. WGBS libraries could also be prepared using transposase-based library construction or tagmentation-WGBS (T-WGBS; also known as Tn5mC-Seq) [20, 21]. This assay uses a hyperactive Tn5 transposase derivative to fragment double-strand DNA and to append methylated adaptors in a single steps. Another method for whole-genome methylation sequencing is post-bisulfite adaptor tagging (PBAT) [22]. This method generates bisulfite sequencing library from bisulfite-converted single-strand DNA with two rounds of random priming with primers containing four random nucleotides on the 3’ end. Generation of sequencing library after bisulfite-treatment reduces the loss of adaptor-ligated DNA as a consequence of DNA damage during bisulfite treatment. A recent developed DNA SMART (Switching Mechanism at 5’ End of RNA Template) technology [23] uses template-switching approach to generate complementary strand DNA from single strand DNA template and directly add adaptors to DNA without adaptor ligation and clean up steps of the regular shortgun library preparation.

Although DNA sequencing cost has dramatically reduced, whole genome sequencing is still difficult to apply for many studies based on human genome and large cohort study. To reduce sequencing cost, selection or enrichment of DNA fragments containing a high level of CpGs is performed before bisulfite sequencing library construction. Reduced representation bisulfite sequencing (RRBS) method has similar procedures to WGBS, but it takes the advantage of methylation-insensitive restriction enzymes such as *Msp*I to fragment and enrich for CpG-rich sequences that are predominantly in CGIs. This allows coverage approximately 10% of total CpGs [24]. Double-digestion such as *Msp*I and *Ape*KI [25] or *Msp*I and *Taq*I increases CpG coverage up to 20% of CpGs. Modification of RRBS methods such as lacer-capture micro-dissection (LCM-RRBS) [26] or multiplexing RRBS (mRRBS) [24] and single cell RRBS (scRRBS) [27] have been successfully used to profile DNA methylation on input as low as 1 ng and in a single cell, respectively. Although RRBS shares similar features of WGBS method, the coverage by RRBS method is limited to only CpG-rich region. The low-CpG density region that exist in enhancer and intronic regions are not adequately covered by this method [25-28]. Other assays have used methylation restriction enzyme sequencing (MRE-seq), methylation immunoprecipitation sequencing (MeDIP-seq) and methyl-CpG-binding domain protein sequencing (MBD-seq) to enrich for methylated DNA fragments. The techniques using methylation-sensitive enzymes could be coupled with sequencing in the MRE-seq protocol [29]. The advantage of MeDIP-seq is it can quantify 5mC level at a large fraction of repeats than other sequencing-based method [29, 30], and it can capture ~90% of total CpG coverage with 17-18 Gb of sequencing [31]. There are several technical caveats related to the methods using methylation enrichment. For instance, they quantify methylation as the relative abundance of 5mC in genomic window of various sizes not in single-base resolution. Copy number variation can cause the bias, so control experiment is required to normalize the difference of copy number at the genomic level.

5mC at specific genomic regions could be selectively quantified by targeted methylation sequencing. These approaches have to be carried out using PCR amplification, ligation capture, bisulfite padlock probe (BSPP) capture, or liquid hybridization capture. The challenge of PCR-based method is being able to multiplex PCR amplification of hundreds to thousand of targets simultaneously without introducing cross-reactions of the PCR primers. Raindance microdroplet PCR technology enables PCR amplification of singleplex in emulsion droplets as high as 20,000 targets [32]. However, DNA input requirement is proportional to the number of targets as each droplet requires multiple copies of genomic template. Another targeted bisulfite sequencing method is ligation capture. This approach enriches DNA targeted fragments by annealing designed oligonucleotides to enzymatically digested DNA followed by ligation with common adaptor for amplification. The two notable ligation capture methods are methylation target capture and ligation (mTACL) [33] and bisulfite patch PCR [34]. In the BSPP method [35, 36], the genomic target regions are captured by padlock probes. A padlock probe contains two short capture sequences that are linked by a common linker sequence. BSPP are designed to be complementary to bisulfite-converted DNA and the CpG(s) to be analyzed are between the two annealing sequences. BSPP method has high flexibility and scalability in selecting target region at various sizes. Hundreds to hundreds of thousands of probes can be pooled into a single capture reaction. The set of 330,000 probes can consistently capture more than 500,000 CpG sites [35]. Liquid hybridization capture capture-based bisulfite sequencing (a), which has been successfully used in exom capture and sequencing, was recently adopted for targeted bisulfite sequencing [37, 38]. For this approach, a sequencing library generated by shotgun library preparation could be enriched by hybridization to the designed biotinylated oligonucleotides, bisulfite converted, and amplified. To allow this method to be applicable for low DNA input, targeted fragments can be enriched from a post-amplified WGBS library with custom biotinylated oligonucleotides complement to post-conversion DNA [39].

Array-based assays have been widely adopted for use in many studies because of theirs features of low costs, ease of use and high throughput. The comprehensive high-throughput arrays for relative methylation (CHARM) [40] and the Illumina Infinium bead chips [41] are the two arrays widely used. CHARM chips do not provide single-base resolution but can be coupled with any methylation enrichment protocol. The flexibility of CHARM array is that users can design custom array for specific purpose. This feature allow for quantification of non-CpG methylation and repetitive regions. The Illumina Infinium 450K BeadChips is a comprehensive array that can interrogate more than 450,000 methylation sites at single-base resolution. A small fraction of non-CpG methylation included in 450K BeadChips. Microarrays have been coupled with MeDIP or MDB (mCIP-chip, MeDIP-chip, MeDIP-on-RepArray, MDB-chip) to specifically target promoters and repeat regions [42-44]. Cross hybridization in array-based assays remains a primary source of bias.

1.3 Padlock probe technology

Padlock probes are single strand DNA molecule with total length approximately 80-100 nucleotides, consisting of two target-complementary sequences of 20-nucleotide long located at both 5'- and 3'-end and connected by a 40-nucleotide long common linker sequences. Padlock probes were firstly reported by Neilson M. in 1994 [45]. Once hybridized to the DNA target, two ends of the target-complementary segments are brought adjacent to each other without the gap between and sealed by DNA ligation resulting in circularized probes. The captured molecules are amplified and sequenced using amplification primers annealing to the sequences on common linker sequences. Several features of padlock probe allowed it to be used in many applications for genomic studies. Capture with padlock probe is highly specific as ligation of the two end of target-complementary segments would occur only when they hybridize to the targeted sequences, and ligation is unlikely to allow for any mismatches at the ligation junction. With this feature, padlock probes of different targets can be pooled to capture in the same reaction with low chance of cross-reactions compared to PCR reaction with many pairs of disjointed primers. Padlock probes could also be designed to uncover the genomic sequences of regions residing in the gap between the two capturing arms. After hybridization, the open gap is filled by DNA polymerase and circularized by DNA ligase. Padlock probes have been adopted in a wide-range of genetic investigations, including genotyping, exome re-sequencing, gene expression studies, and *in situ* hybridization [46-54]. Bisulfite padlock probe was modified from padlock probe as the target-complementary segment or capturing arms were designed to be complementary to bisulfite-converted sequences. The very first generation of bisulfite padlock probes

(BSPP) [36] have been used for chromosome-wide DNA methylation quantification of ~66,000 CpG sites within 2,020 CpG islands on human chromosome 12 and 20 using ~33,000 BSPPs [36]. BSPPs have also been used to capture DNA methylation on promoters and gene bodies using pools of ~10,000 BSPPs [55]. The most recent improvement to BSPPs [35] have increased coverage of the human genome with ~330,000 probes that cover genomic locations known to contain differentially methylated regions (DMR) or differentially methylated sites (DMSs), transcriptional repressor CTCF binding sites, DNase I-hypersensitive regions, all micro RNA genes and all promoters.

In the next several chapters, I will cover the development of targeted methylation analysis method using bisulfite padlock probe (BSPP) to characterize DNA methylation status at informative loci of human genome. Chapter 3 focuses on the extensive validation of the performance of BSPP for locus-specific DNA methylation analysis. In chapter 4, I apply the DMR330k probe set on 96 samples from 22 nuclear pedigrees to investigate the regulation of variability of DNA methylation by genetic polymorphisms.

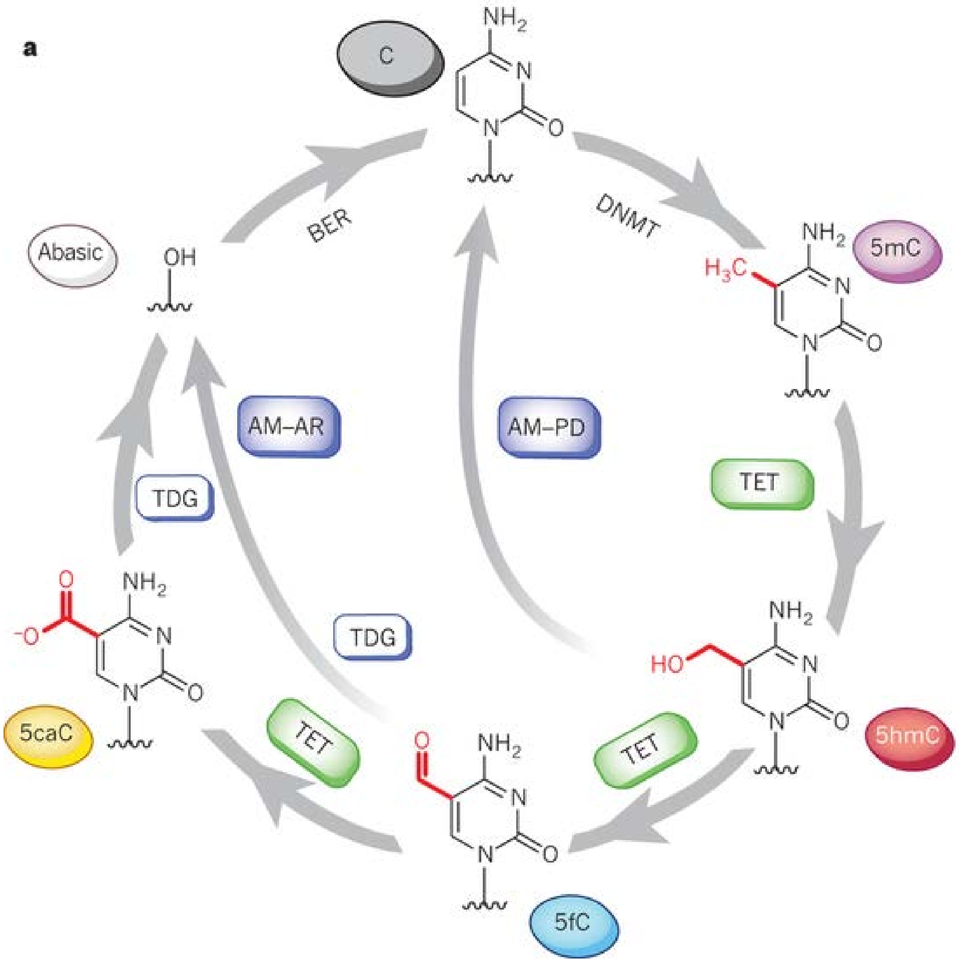


Figure 1.1 A biochemically pathway for modification of cytosine. 5-methylcytosine (5mC) is methylated by DNA methyltransferase (DNMT) enzymes, and can be oxidized sequentially to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by ten-eleven translocation (TET) family enzymes. (Adapted from Kohli RM and Zhang Y, Nature 2013, [56])

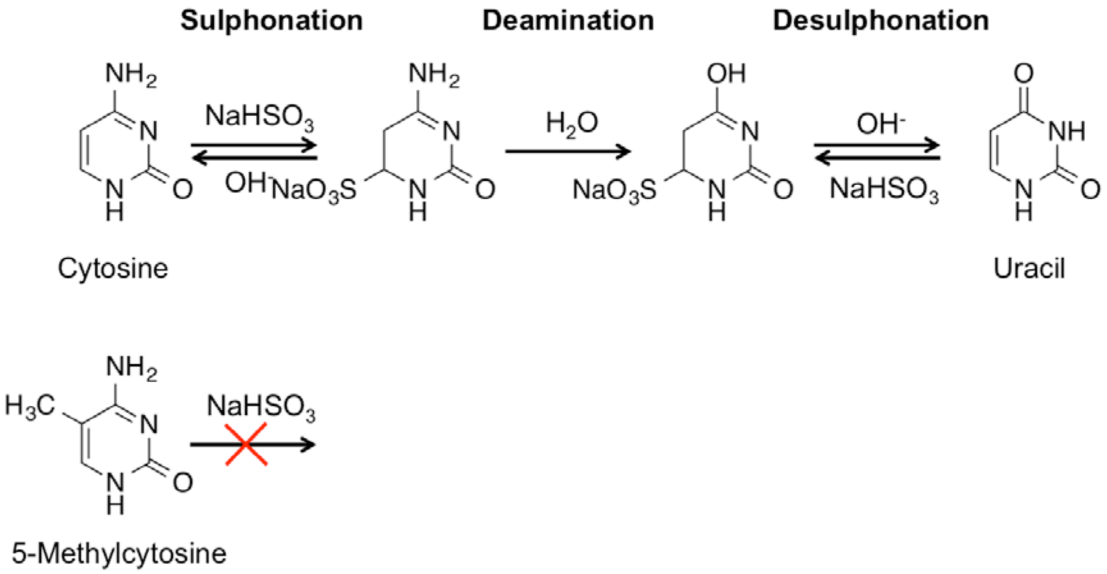


Figure 1.2 Bisulfite conversion procedures. Treatment DNA with sodium bisulfite converted unmethylated cytosine into uracil by the sequential reactions, sulphonation, deamination and desulphonation. 5mC or 5hmC resist to bisulfite treatment and remain intact.

Table 1.1 Overview of 5-methylcytosine (5mC) quantification methods. (Adapted from Plongthongkum N, Diep DH, and Zhang K, 2014, [57])

DNA modification	**Measurement**	**Non-targeted enrichment**	**Targeted enrichment**	**Whole genome**	**Arrays**
5mC	Absolute (single base)	RRBS, mRRBS, LCM-RRBS or scRRBS	Microdroplet bisulfite PCR, Bisulfite Patch PCR, mTACL, BSPP, LHC-BS (pre- and post-conversion) or RSMA	WGBS, T-WGBS or PBAT	Infinium BeadChip
	Relative (peak)	MRE-seq, MeDIP-seq, MBD-seq or MethylCap-seq			CHARM or MeKL-ChIP

1.4 References

1. Arand J, Spieler D, Karius T, Branco MR, Meilinger D, Meissner A, Jenuwein T, Xu GL, Leonhardt H, Wolf V and Walter J (2012) In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases. Plos Genetics 8(6).

2. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD and Ren B (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet 45:1198-1206.

3. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B and Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315-322.

4. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM and Ecker JR (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471:68-73.

5. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao A, Esteller M, He C, Haghighi FG, Sejnowski TJ, Behrens MM and Ecker JR (2013) Global Epigenomic Reconfiguration During Mammalian Brain Development. Science.

6. Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T, Gnirke A and Meissner A (2011) Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. PLoS Genet 7(12).

7. Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, Wang H, Hannah R, Kim S-B, Yang L, Ko M, Chen R, Gottgens B, Lee J-S, Gunaratne P, Godley LA, Darlington GJ, Rao A, Li W and Goodell MA (2014) Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat Genet 46:17-23.

8. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, Nimwegen Ev, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK and Schübeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature.

9. Xie W, Schultz Matthew D, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker John W, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee Ah Y, Swanson Scott A, Zhang J, Zhu Y, Kim A, Nery Joseph R, Urich Mark A, Kuan S, Yen C-a, Klugman S, Yu P, Suknuntha K, Propson Nicholas E, Chen H, Edsall Lee E,

Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung W-Y, Chi Neil C, Antosiewicz-Bourget Jessica E, Slukvin I, Stewart R, Zhang Michael Q, Wang W, Thomson James A, Ecker Joseph R and Ren B (2013) Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell 153:1134-1148.

10. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL and Lander ES A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125:315-326.

11. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 13:484-492.

12. Smith ZD and Meissner A (2013) DNA methylation: roles in mammalian development. Nature reviews. Genetics 14:204-220.

13. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, Gascard P, Sigaroudinia M, Tlsty TD, Kadlecek T, Weiss A, O'Geen H, Farnham PJ, Madden PAF, Mungall AJ, Tam A, Kamoh B, Cho S, Moore R, Hirst M, Marra MA, Costello JF and Wang T (2013) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet 45:836-841.

14. Yu M, Hon GC, Szulwach KE, Song C-X, Jin P, Ren B and He C (2012) Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nature Protocols 7:2159-2170.

15. Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P, Tahiliani M, Daley GQ, Liu XS, Ecker JR, Milos PM, Agarwal S and Rao A (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature 473:394-397.

16. Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung H-L, Zhang K and Zhang Y (2013) Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. Cell.

17. Szulwach KE, Li X, Li Y, Song C-X, Wu H, Dai Q, Irier H, Upadhyay AK, Gearing M, Levey AI, Vasanthakumar A, Godley LA, Chang Q, Cheng X, He C and Jin P (2011) 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nature Neuroscience 14:1607-1616.

18. Iurlaro M, Ficz G, Oxley D, Raiber E-A, Bachman M, Booth MJ, Andrews S, Balasubramanian S and Reik W (2013) A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biology 14(10).

19. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PWTC, Bauer C, Münzel M, Wagner M, Müller M, Khan F, Eberl HC, Mensinga A, Brinkman AB, Lephikov K, Müller U, Walter J, Boelens R, van Ingen H, Leonhardt H, Carell T and Vermeulen M (2013) Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. Cell 152:1146-1159.

20. Adey A and Shendure J (2012) Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. Genome Res. 22:1139-1143.

21. Wang Q (2013) Tagmentation-based whole-genome bisulfite sequencing. Nature Protoc. 8:2022-2032.

22. Miura F (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. 40:e136.

23. Ramskold D, Luo SJ, Wang YC, Li R, Deng QL, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP and Sandberg R (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology 30:777-782.

24. Boyle P (2012) Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol. 13:R92.

25. Wang J (2013) Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. BMC Genomics 14:11.

26. Schillebeeckx M (2013) Laser capture microdissection-reduced representation bisulfite sequencing (LCM-RRBS) maps changes in DNA methylation associated with gonadectomy-induced adrenocortical neoplasia in the mouse. Nucleic Acids Res. 41:e116.

27. Guo H (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 23:2126-2135.

28. Ruiz S, Diep D, Gore A, Panopoulos AD, Montserrat N, Plongthongkum N, Kumar S, Fung HL, Giorgetti A, Bilic J, Batchelder EM, Zaehres H, Kan NG, Scholer HR, Mercola M, Zhang K and Izpisua Belmonte JC (2012) Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. Proc Natl Acad Sci U S A 109:16196-201.

29. Maunakea AK (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253-257.

30. Shen L (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell 153:692-706.

31. Clark C (2012) A comparison of the whole genome approach of MeDIP-seq to the targeted approach of the infinium HumanMethylation450 BeadChip[reg] for methylome profiling. PLoS ONE 7:e50233.

32. Komori HK (2011) Application of microdroplet PCR for large-scale targeted bisulfite sequencing. Genome Res. 21:1738-1745.

33. Nautiyal S (2010) High-throughput method for analyzing methylation of CpGs in targeted genomic regions. Proc. Natl Acad. Sci. USA 107:12587-12592.

34. Varley KE and Mitra RD (2010) Bisulfite Patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. Genome Res. 20:1279-1287.

35. Diep D (2012) Library-free methylation sequencing with bisulfite padlock probes. Nature Methods 9:270-272.

36. Deng J (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nature Biotech. 27:353-360.

37. Lee EJ (2011) Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. Nucleic Acids Res. 39:e127.

38. Wang J (2011) High resolution profiling of human exon methylation by liquid hybridization capture-based bisulfite sequencing. BMC Genomics 12:597.

39. Ivanov M (2013) In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. Nucleic Acids Res. 41:e72.

40. Irizarry RA (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res. 18:780-790.

41. Bibikova M (2011) High density DNA methylation array with single CpG site resolution. Genomics 98:288-295.

42. Yalcin A (2013) MeDIP coupled with a promoter tiling array as a platform to investigate global DNA methylation patterns in AML cells. Leukemia Res. 37:102-111.

43. Gilson E and Horard B (2012) Comprehensive DNA methylation profiling of human repetitive DNA elements using an MeDIP-on-RepArray assay. Methods Mol. Biol. 859:267-291.

44. Zhang X (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell 126:1189-1201.

45. Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary BP and Landegren U (1994) Padlock Probes - Circularizing Oligonucleotides for Localized DNA Detection. Science 265:2085-2088.

46. Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon J-K, Rosenbaum AM, Zaranek AW, LeProust E, Sunyaev SR and Church GM (2009) Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. Genome Research 19:1606-1615.

47. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, LeProust E, Zhang K, Gao Y and Church GM (2009) Genome-Wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing. Science 324:1210-1213.

48. Lee JH, Park IH, Gao Y, Li JB, Li Z, Daley GQ, Zhang K and Church GM (2009) A Robust Approach to Identifying Tissue-Specific Gene Expression Regulatory Variants Using Personalized Human Induced Pluripotent Stem Cells. Plos Genetics 5(11).

49. Wang H, Chattopadhyay A, Li Z, Daines B, Li YM, Gao CX, Gibbs R, Zhang K and Chen R (2010) Rapid identification of heterozygous mutations in Drosophila melanogaster using genomic capture sequencing. Genome Research 20:981-988.

50. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM and Shendure J (2007) Multiplex amplification of large sets of human exons. Nat Meth 4:931-936.

51. Turner EH, Lee C, Ng SB, Nickerson DA and Shendure J (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Meth 6:315-316.

52. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM, Eggan K and Church GM (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nature Methods 6:613-U90.

53. Nilsson M, Krejci K, Koch J, Kwiatkowski M, Gustavsson P and Landegren U (1997) Padlock probes reveal single-nucleotide differences, parent of origin and in situ distribution of centromeric sequences in human chromosomes 13 and 21. Nat Genet 16:252-255.

54. Larsson C, Koch J, Nygren A, Janssen G, Raap AK, Landegren U and Nilsson M (2004) In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. Nature Methods 1:227-232.

55. Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ and Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nature Biotechnology 27:361-368.

56. Kohli RM and Zhang Y (2013) TET enzymes, TDG and the dynamics of DNA demethylation. Nature 502:472-479.

57. Plongthongkum N, Diep DH and Zhang K (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet 15:647-661.

CHAPTER 2 LIBRARY-FREE METHYLATION SEQUENCING WITH BISULFITE PADLOCK PROBES

2.1 Abstract

Targeted quantification of DNA methylation allows for interrogation of the most informative loci across many samples quickly and cost-effectively. Here we report improved bisulfite padlock probes (BSPPs) with a design algorithm to generate efficient padlock probes, a library-free protocol that dramatically reduces sample-preparation cost and time and is compatible with automation, and an efficient bioinformatics pipeline to accurately obtain both methylation levels and genotypes from sequencing of bisulfite-converted DNA.

2.2 Introduction

We have previously developed bisulfite padlock probes for the specific and parallel digital quantification of DNA methylation [1]. Recently, we enhanced BSPPs for improved flexibility and multiplexing capability. These improvements have contributed to recent findings in mouse and human pluripotent stem cells [2-5].

First, target selection and probe design is crucial for BSPPs. To aid in the design of efficient padlock probes for bisulfite analysis, we developed a program called ppDesigner. It accepts as input the genome of any organism, a user's list of arbitrary targets and user-desired probe constraints matching requirements of the experimental protocol. It *in silico* 'bisulfite-converts' the genome (that is, it changes all cytosines to thymines) and outputs padlock probes to cover the chosen targets while avoiding CpGs on the capturing arms that could be methylated and not converted to be recognized as thymine. ppDesigner uses a back-propagation neural network to predict probe efficiency (Figure 2.1). We had previously trained this network using data from probes for exomic targets [6] based on seven properties. Using bisulfite capture data from the first BSPPs [1], we refined the network with two additional factors. ppDesigner can explain ~50% of the variance in capturing efficiency for genomic DNA and ~20% of the variance in capturing efficiency for bisulfite-converted DNA; additional variation could be due to factors such as variability in oligonucleotide synthesis and sample DNA quality. ppDesigner is extremely flexible and has been used to design a variety of genomic and bisulfite probes for *Homo sapiens* [2, 3]*,* *Mus musculus* [4] and *Drosophila melanogaster* [7].

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 [8] 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 2.1). 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 (Table 2.1 and 2.1). For large sample sizes, the library preparation cost (including probes) with our protocol was comparable to that of the restricted-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. Restricted-representation bisulfite sequencing is more cost-effective than BSPPs, but the former lacks BSPPs' flexibility in selecting specific sites or regions.

Another bottleneck in sequencing of bisulfite-converted DNA is a lack of computational tools to efficiently analyze sequencing data generated from hundreds of samples. To overcome this issue, we developed an analysis pipeline for read mapping and methylation quantification, called bisReadMapper (Figure 2.3). In previous padlock probe studies, reads had been mapped only against target regions owing to the computational requirements of sequence alignment [1]. In contrast, we designed bisReadMapper to map to the full genome sequence, allowing processing of data from both targeted and whole-genome sequencing of bisulfite-converted DNA. bisReadMapper also 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- and unidirectional bisulfite libraries in an unbiased manner. Another feature is the capability to call single-nucleotide polymorphisms (SNPs) from sequences of bisulfite-converted DNA; this feature not only allows for analysis of allele-specific methylation [9] but also allows accurate sample tracking in large-scale experiments. Finally, bisReadMapper can call methylation levels at both CpG and non-CpG sites.

2.3 Materials and Methods

2.3.1 Bisulfite padlock probe production

2.3.1.1 Oligonucleotides from Agilent (Long oligonucleotides of 150 nucleotides (nt))

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 µg of the purified amplicons were digested with 50 units of Lambda Exonuclease (5U/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 (1U/µl, NEB) at 37 ºC for 1 h. The digested DNAs were annealed to 5.88 uM 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 DpnII (10U/ul, 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/DpnII 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).

2.3.1.2 Oligonucleotide from LC Sciences (Short version of oligonucleotides of 100 nucleotides (nt))

The oligonucleotides (100nt) were synthesized by the programmable microarray platform (LC Sciences) and released to form the mix of 4,000 oligoucleotides. The oligonuclotides were amplified by two-step PCR in 200 µl reacton with 1nM template oligonucleotides, 400 nM each of eMIP_CA1_F primer and eMIP_CA1_R primer (supplementary Table 3) and 100 µl of KAPA SYBG fast Universal qPCRMaster 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 amplified amplicons were purified with Qiaquick PCR purification columns and re-amplified by PCR in 32 reactions (100 µl each) with 0.02nM first round amplified amplicons, 400nM each of eMIP_CA1_F primer and eMIP_CA1_R primer and 50 µl of KAPA SYBG fast Universal qPCRMaster 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 amplified 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 (100U/µl, NEB) at 37 ºC for 1 h in NEBuffer 2. The enzyme was heat inactivated at 80 ºC for 20 min. Then the reaction was incubated with 100 units of Nb.BrsDI (10 1U/µl, NEB) at 65 ºC for 1 h. The nicked–DNAs were purified by

Qiaquick PCR purification column. The probe size approximately 70 nucleotides were purified in 6% denaturing PAGE (6% TB-urea 2D gel).

2.3.2 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, 200ng of bisulfite converted gDNA, and 4.2 nM oligo suppressor were mixed in 25ul 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. 2.5ul of SLN mix (100 uM dNTP, 2U/ul AmpliTaq Stoffel Fragment (ABI) and 0.5 U/ul 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.

2.3.3 Capture circles amplification (Library-free BSPP protocol)

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

2.3.3.1 Capture circles amplification (Probe from LC Sciences)

10 µl circularized DNA was amplified in 100 µl reaction with 200nM each of CP-2-FA primer and CP-2-RA primer (Supplementary Table 3), and 50 µl KAPA SYBG fast Universal qPCRMaster 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, and 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 amplified amplicons with the corresponding expected size approximately 260 bp were purified with 6% PAGE (6% 5-well gel, Invitrogen) and resuspended with 12 µl of TE buffer. 30% of the gel-purified amplicons were re-amplified and barcoded in 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 SYBG fast Universal qPCRMaster 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.

2.3.3.2 Capture circles amplification (N2-adapter BSPP protocol)

The captured DNA library was amplified as previously described (1). Briefly, 8-10 µl of capture reaction mix was amplified by PCR in 100 µl reactions with 400 nM each of AmpF6.3NH2 primer and AmpR6.3NH2 primer (Supplementary Table 2), 0.4x SYBR Green I (Invitrogen), and 50 µl Phusion High-Fidelity 2x Master Mix (NEB) at 98 ºC for 30 seconds, 13-20 cycles of 98 ºC for 10 seconds; 58 ºC for 20 seconds; and 72 ºC for 20 seconds, and 72 ºC for 3 minutes. Amplicons were purifed with 6% TBE PAGE gel (Invitrogen) and sequenced on Illumina Genome Analyzer IIx.

2.3.4 Bisulfite sequencing library construction

300 ng of each capture amplicon was digested at 37 ºC for one hour in 5 units of MmeI (NEB) in NEBuffer 4 and 75 µM S-Adenosylmethionine (NEB). Digested

products were purified with one Qiaquick PCR purification column each. Adaptor ligation was carried out with 20 μL of digested product, 0.5 μM adaptor mix per 1 ng digested product, 1x Quick Ligase buffer (NEB), and 2 μl of Quick T4 DNA ligase (NEB), and incubated at room temperature for 15 minutes. The adaptor mix was prepared ahead of time by mixing 20 μM PE_t_N2 and 20 uM PE_b_A (Supplementary Table 2) in equimolar ratio and then incubating at 94 ºC for 3 min; temperature was gradually lowered at 0.1 ºC/s to 20 ºC in a thermocycler. Adaptor ligated products were purified with 0.7x volume AMPure beads (Agencourt) and purified using Qiagen Qiaquick columns into 40 μL of elution buffer. A quarter of the eluted DNA for each sample was amplified and barcoded in 100 μl reactions with 200 nM PCR_F (Supplementary Table 2) and 200 nM of barcoded PCR_R.N2.IndX primers (Supplementary Table 3), 0.2x SYBR Green I, and 50 μl Phusion High-Fidelity 2x Master Mix (NEB), at 98 ºC for 30 s, cycled 9-12 times at 98 ºC for 10 s; 64 ºC for 20 s; and 72 ºC for 30 s, and finally 72 ºC for 2 min. Amplicons were purifed with 6% TBE PAGE gel (Invitrogen) and sequenced on Illumina Genome Analyzer IIx.

2.3.5 Bisulfite read mapping and data analysis

Bisulfite converted data was processed as previously described. 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. Methylation frequencies and SNPs were called using a method described previously.

2.4 Results

To test our assay, we generated a genome-scale probe set based on our previous results and new information about differential methylation [1, 10-12]. We targeted our new design for evaluation of methylation at genomic locations known to contain differentially methylated regions or differentially methylated sites (DMSs) [10-13], 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, 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 [1] (Figure 2.4). To improve uniformity, we normalized the experimental capturing performance of these probes using subsetting and suppressor oligonucleotides as described previously [1]. We could characterize roughly 500,000 CpG sites with ~4 gigabases of sequencing reads, and additional sites became callable with deeper sequencing (Figure 2.5 and 2.6).

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 [14] 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 2.7).

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, (Figure 2.8a,b). 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[12](Pearson's correlation coefficient R = 0.95, (Figure 2.9).

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 2.10). 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 (Figure 2.8c-e). 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.

Our library-free BSPP method is flexible for different study designs. 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. 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 cells [15]. 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 2.11), 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 (Figure 2.12) 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.

This method can be implemented to aid in identifying the effects of DNA methylation in any organism by using the computational tools at http://genome-tech.ucsd.edu/public/Gen2_BSPP/.

Chapter 2, in full, is a reprint of the material as it appears in Nature Methods 2012 Vol. 9. Dinh Diep, Nongluk Plongthongkum, Athurva Gore, Ho-Lim Fung, Robert Shoemaker and Kun Zhang. "Library-free methylation sequencing with bisulfite padlock probe." Nature Methods 9(3), (2012): 270-272. The dissertation author was the primary investigator and co-author of this paper.

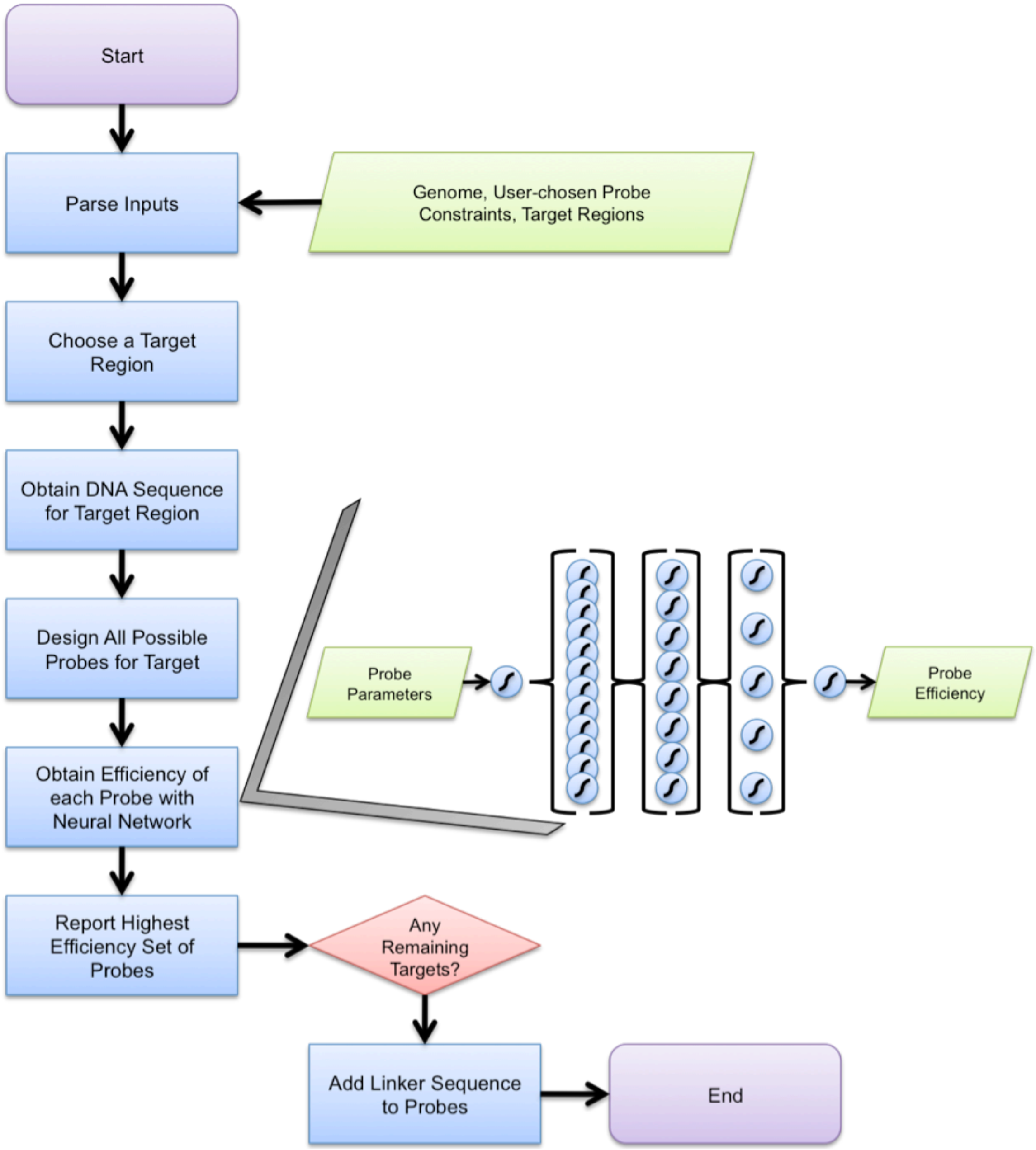

Start
Parse Inputs
Genome, User-chosen Probe Constraints, Target Regions
Choose a Target Region
Obtain DNA Sequence for Target Region
Design All Possible Probes for Target
Obtain Efficiency of each Probe with Neural Network
Probe Parameters
Probe Efficiency
Report Highest Efficiency Set of Probes
Any Remaining Targets?
Add Linker Sequence to Probes
End

Figure 2.1 Schematic for the probe design software (ppDesigner). The neural network model utilizes the target length, target GC content, binding arm melting temperature, binding arm length, local single-stranded folding energy of the target, and the dinucleotides present at the extension site and ligation site during probe capture. Example probes can be found in Figure 2.11.

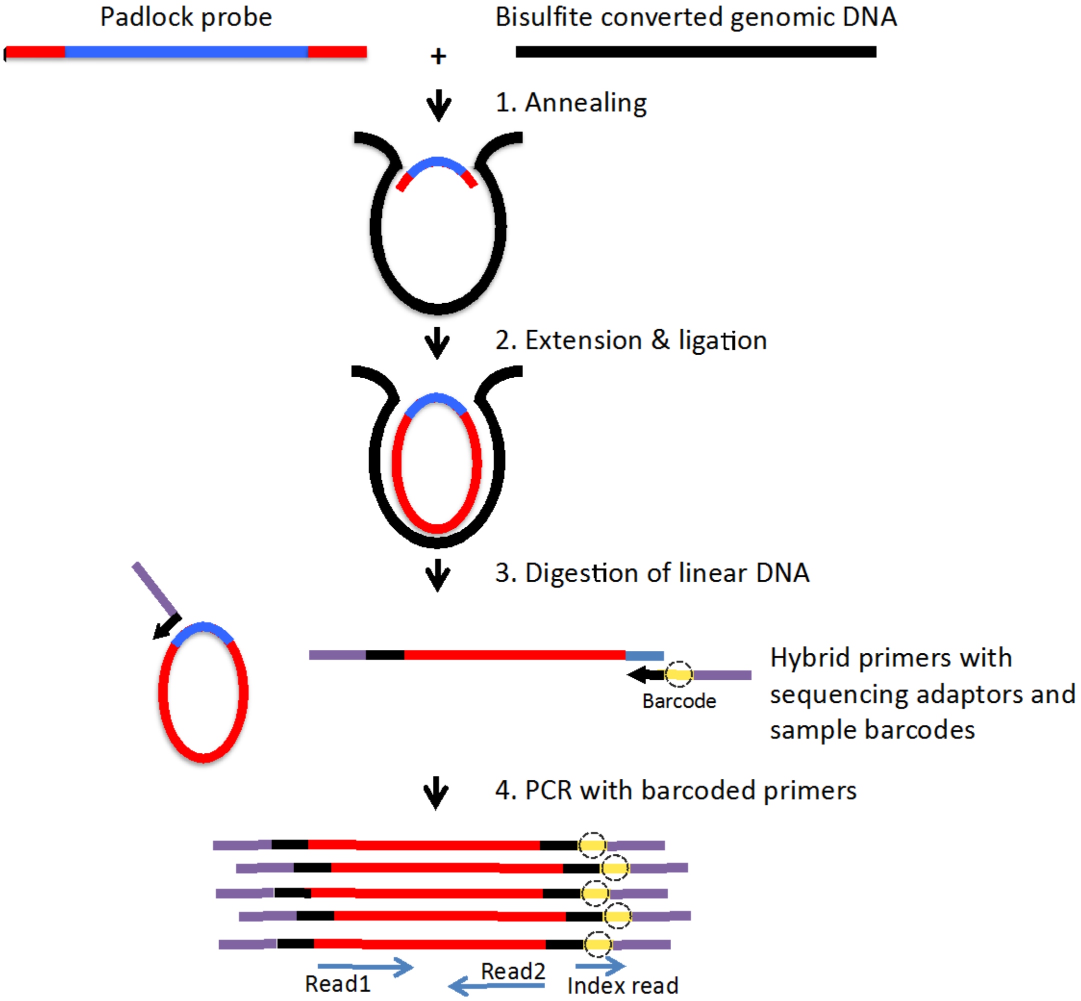

Padlock probe
Bisulfite converted genomic DNA
+
1. Annealing
2. Extension & ligation
3. Digestion of linear DNA
Barcode
Hybrid primers with sequencing adaptors and sample barcodes
4. PCR with barcoded primers
Read1
Read2
Index read

Figure 2.2 Library-free BSPP protocol. Each padlock probe has a common linker sequence flanked by two target-specific capturing arms (red) that anneal to bisulfite converted genomic DNA. The 3' end is extended and ligated with the 5' end to form circularized DNA. After removal of linear DNA, all circularized captured targets are PCR-amplified with barcoded primers and directly sequenced with an Illumina sequencing platform (GA II(x) or HiSeq). Amplicon size is 363bp, which includes captured target (180bp), capturing arms (55bp), and amplification primers and adapters (128bp). The inserts can be read through with paired-end 120-bp sequencing reads.

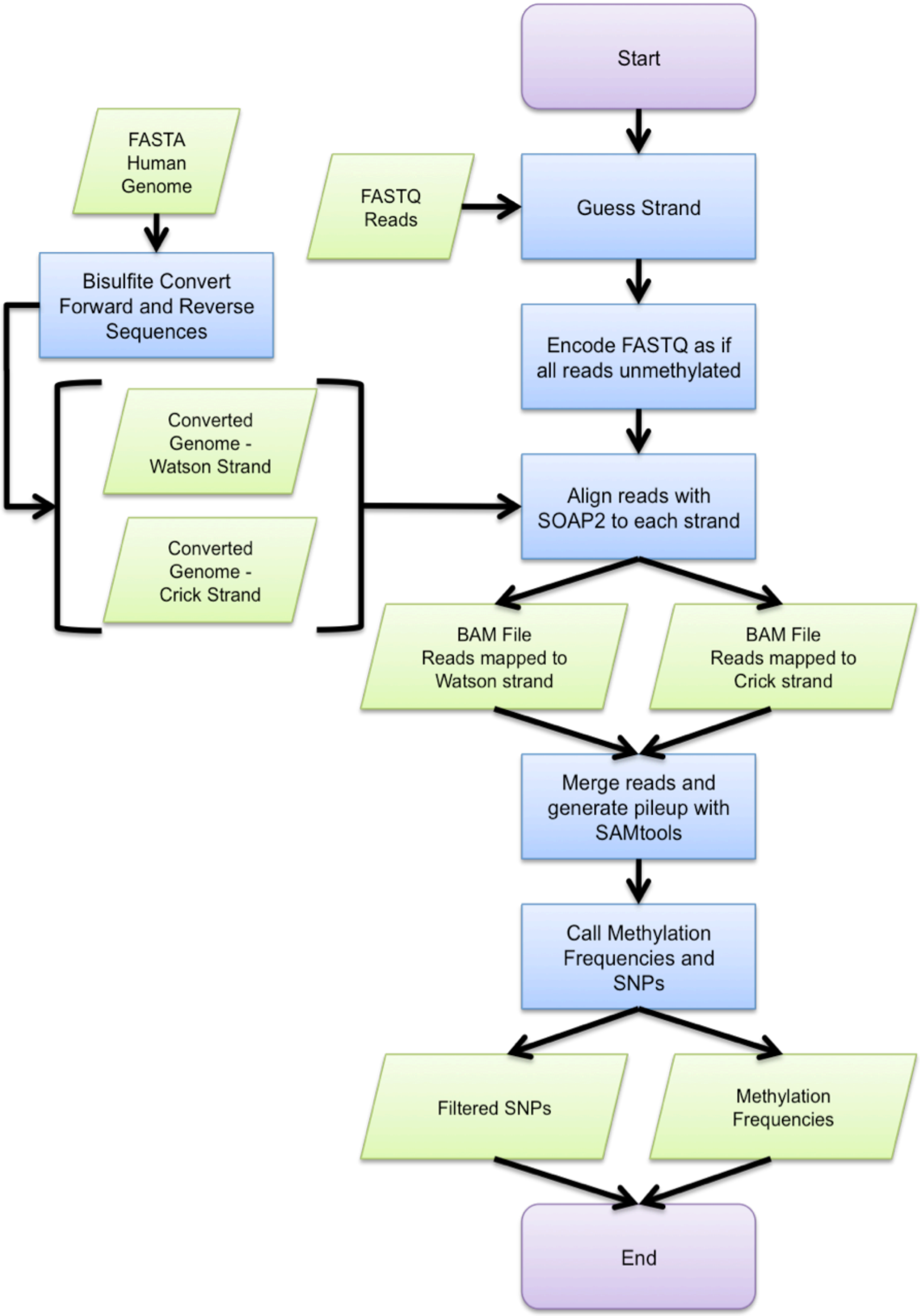

Start
FASTA Human Genome
FASTQ Reads
Guess Strand
Bisulfite Convert Forward and Reverse Sequences
Encode FASTQ as if all reads unmethylated
Converted Genome - Watson Strand
Converted Genome - Crick Strand
Align reads with SOAP2 to each strand
BAM File Reads mapped to Watson strand
BAM File Reads mapped to Crick strand
Merge reads and generate pileup with SAMtools
Call Methylation Frequencies and SNPs
Filtered SNPs
Methylation Frequencies
End

Figure 2.3 Schematic for bisulfite sequencing data analysis pipeline (bisReadMapper)

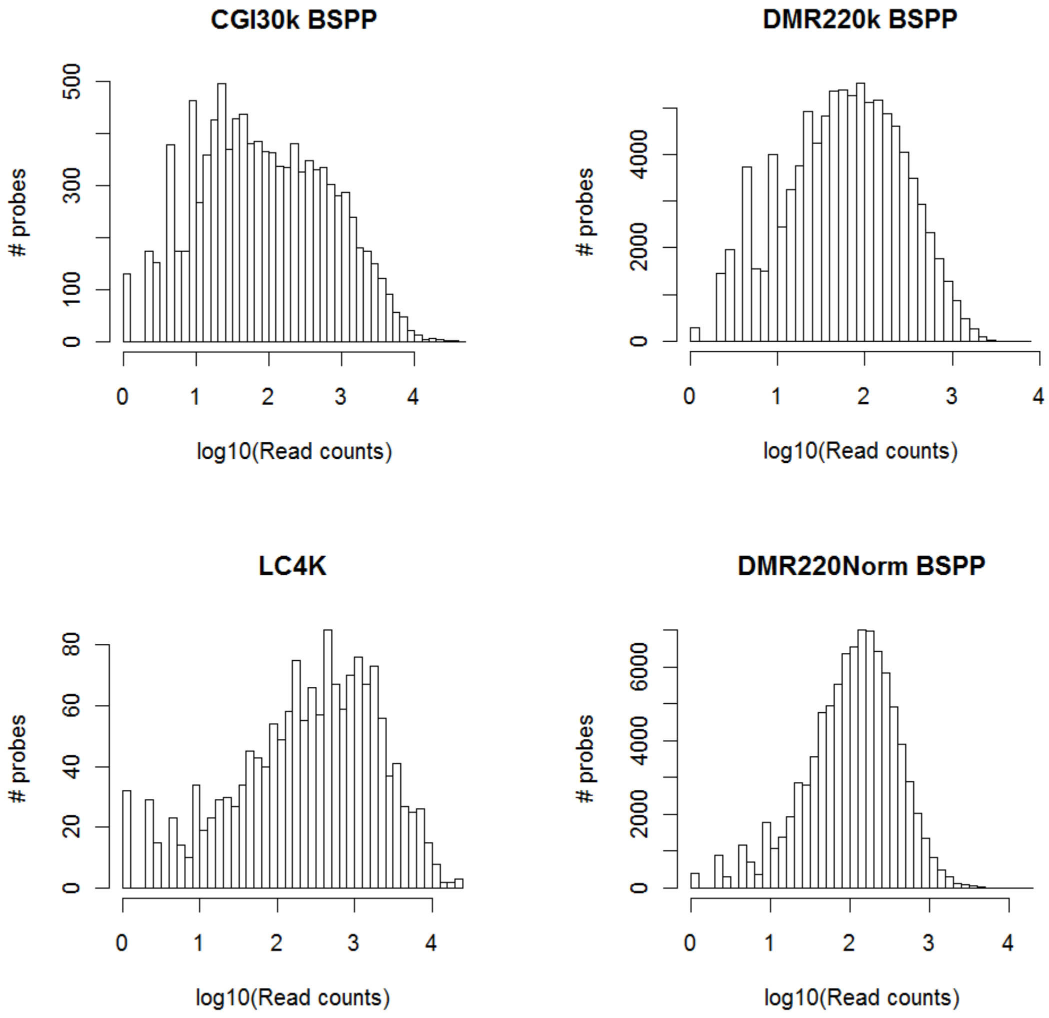


Figure 2.4 Comparison of probe capture efficiencies between the DMR220K, LC4K probe sets and the previously published CGI30K set. The first three plots were generated from data without subsetting or suppressor oligos to allow for a direct comparison of probe design.

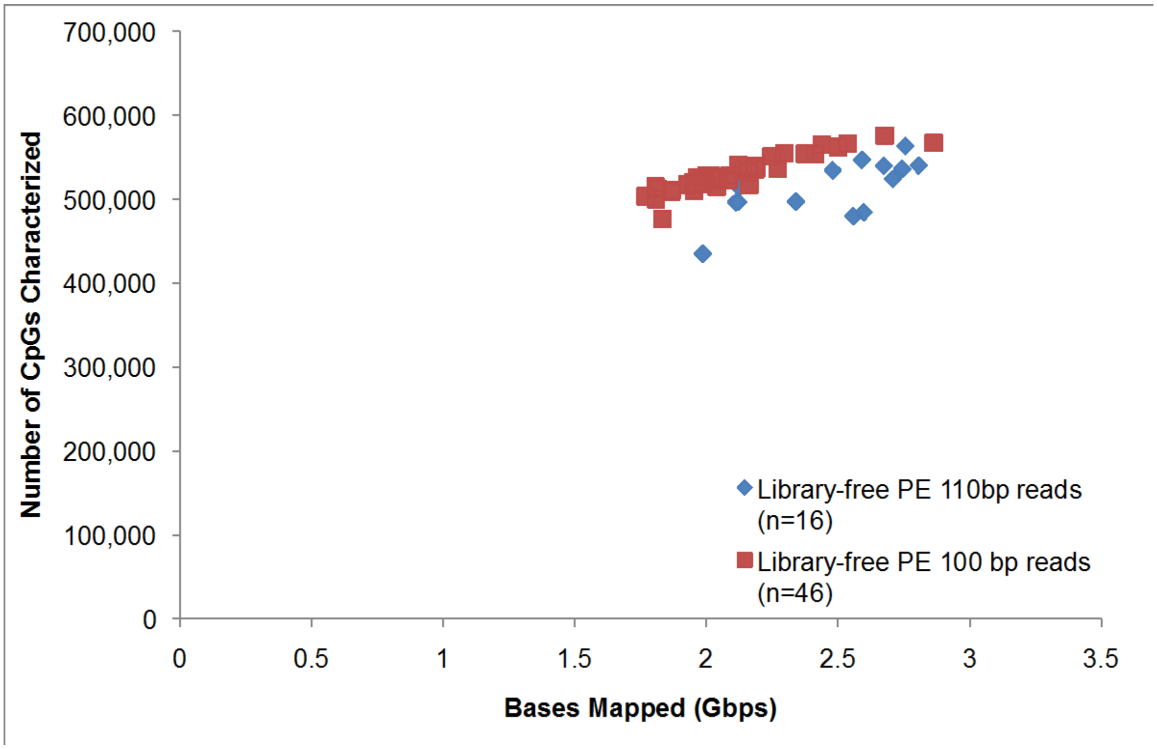


Figure 2.5 Scatter plot of number of characterized CpG sites versus mappable sequencing data for the DMR330K probe set. Variability in sequencing quality of individual sequencing runs is responsible for the different number of CpG sites characterized with similar sequencing effort.

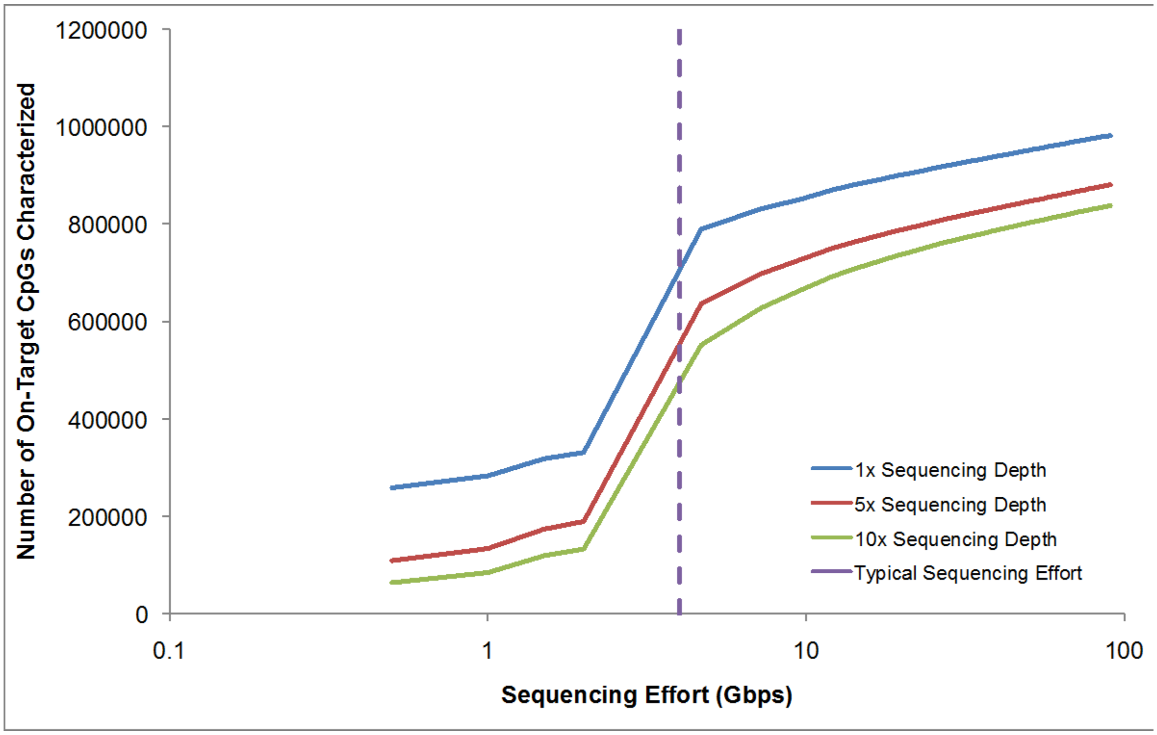


Figure 2.6 Number of CpG sites called per sample as a function of sequencing effort. The horizontal dash line represents 4Gbps of sequences per library that we routinely generate.

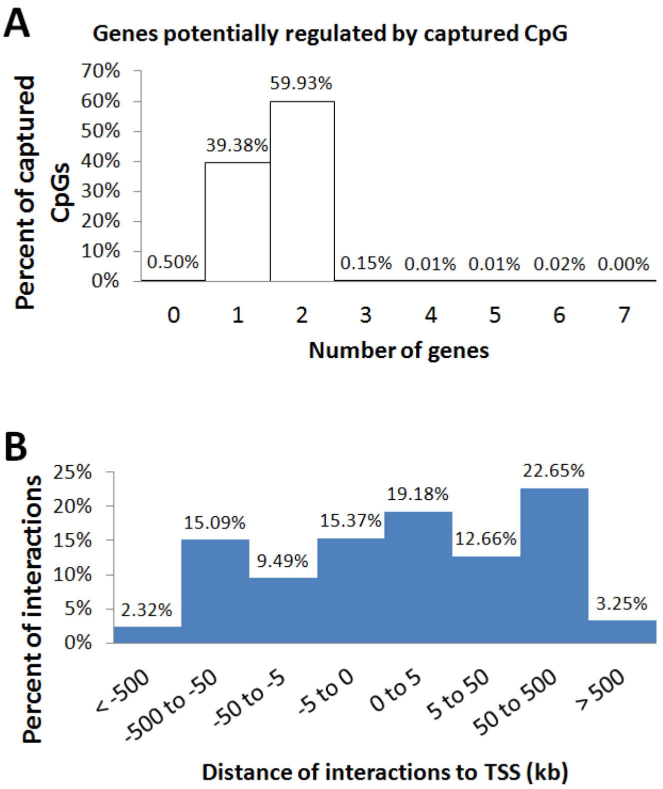


Figure 2.7 Captured CpG sites were tested for potential regulatory interactions with genes by GREAT (http://great.stanford.edu). (A) Most CpG sites were interacting with 1-2 genes. (B) Distance of CpG sites to the transcriptional start sites (TSS) of the predicted regulating genes.

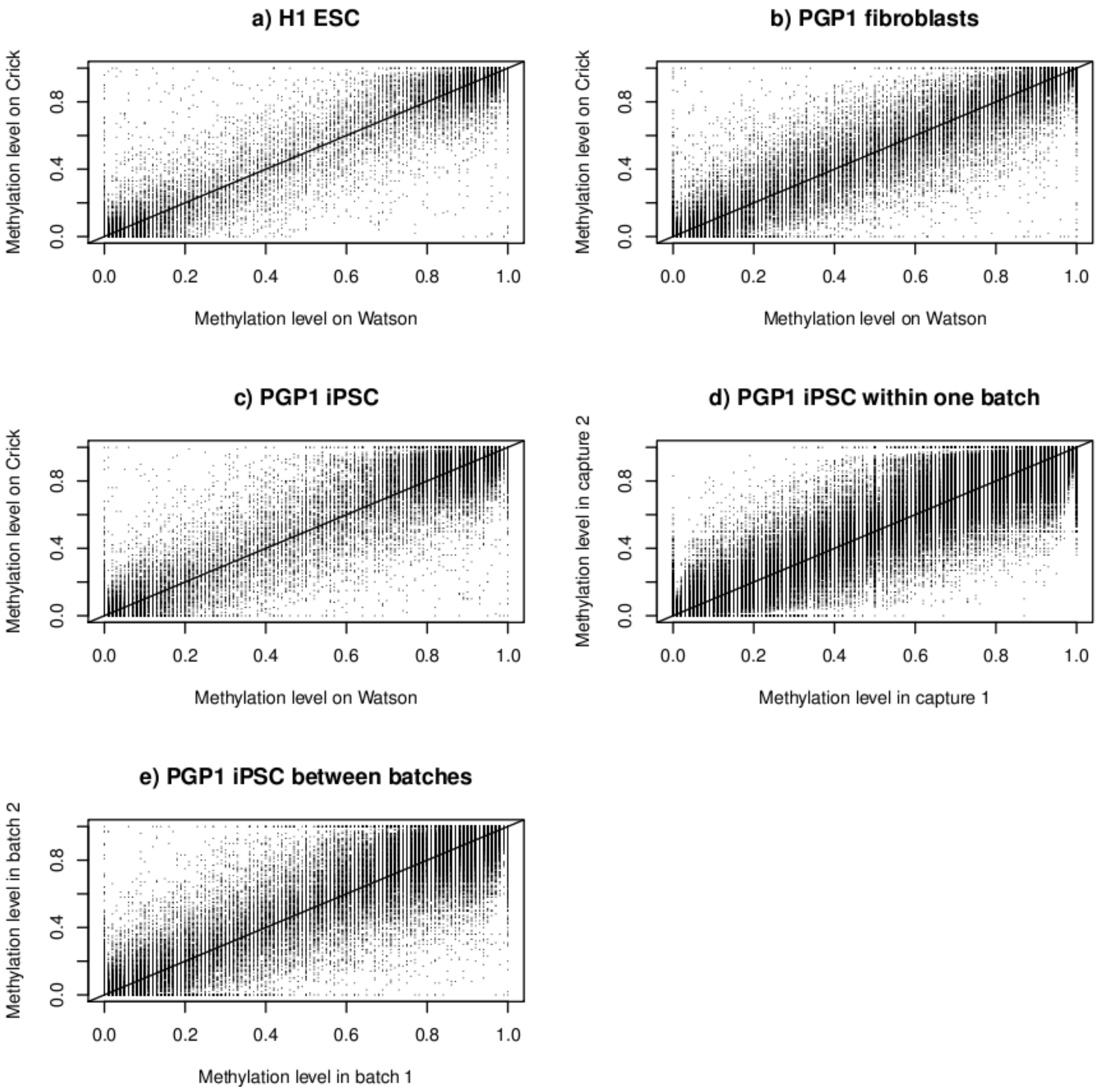


Figure 2.8 Accuracy of digital quantification by BSPP. (a,b) Within batch and between batch comparison of the methylation levels obtained at 10x depth from multiple capture reactions of the same sample (PGP1iPS). The Pearson's correlation coefficient R for within one batch is 0.98 (N=405,508), and for different batch is 0.97 (N=117,186). (c,d,e) Within sample comparison of methylation levels obtained from different probes capturing the same CpG site on different strands at 10x depth within one capture reaction. The Pearson's correlation coefficient R was 0.96 (N=44,361), 0.96 (N=55,965), and 0.97 (N=29,884) for PGP1iPS, PGP1F, and H1 respectively.

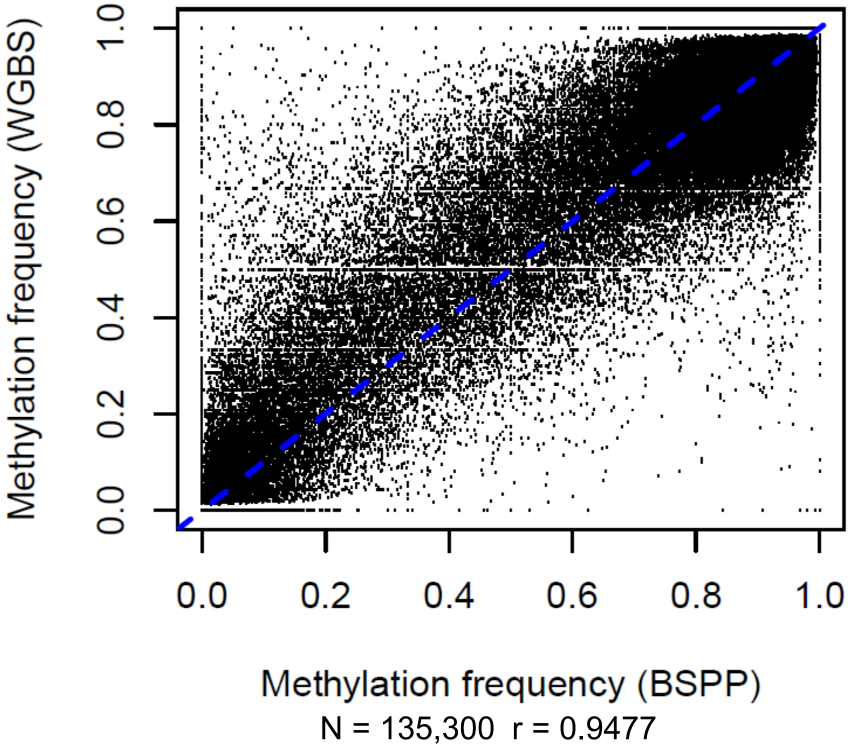


Figure 2.9 Comparison between BSPP and whole genome bisulfite sequencing (WGBS). We compared two H1 ESC datasets, using sites with at least 10x read depth in each. The Pearson's correlation coefficient R was 0.9477 (N=135,300). (Note that the sequencing experiments were performed on separate cultures of H1 from two different labs.)

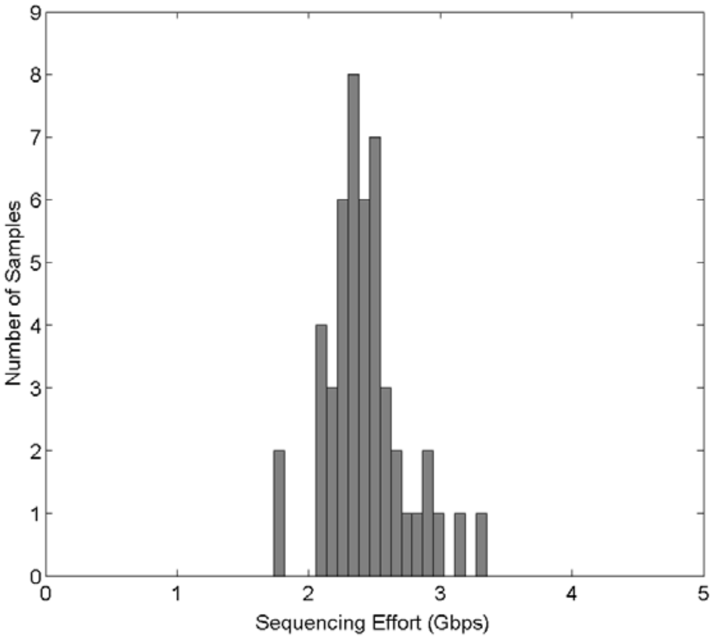


Figure 2.10 Variation in amount of sequencing data obtained per sample in a multiplexed BSPP capture experiment. 48 whole blood samples were captured and sequenced in one batch using the library-free BSPP method. There is little variation between samples in the amount of generated sequencing data.

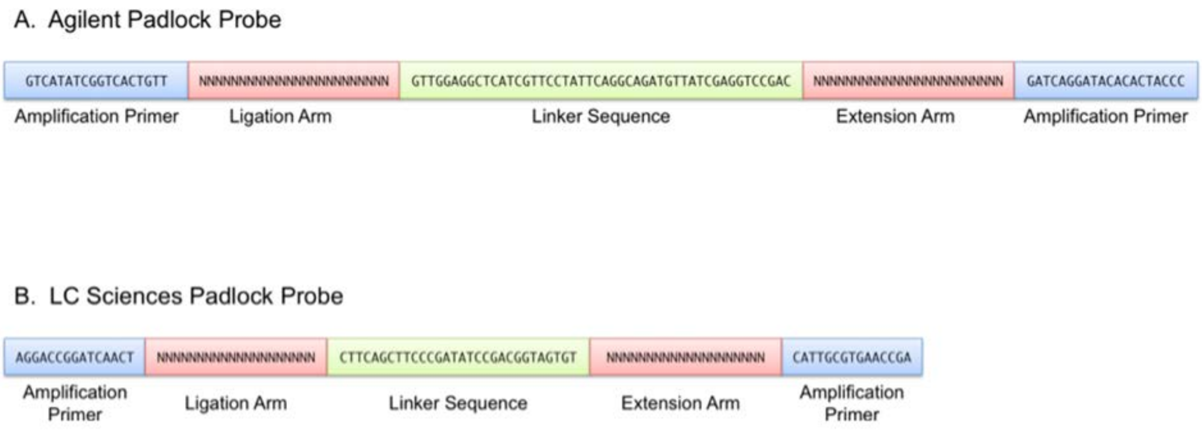


Figure 2.11 Example padlock probes ordered from a) Agilent’s oligonucleotide synthesis service and b) LC Sciences’ oligonucleotide synthesis service.

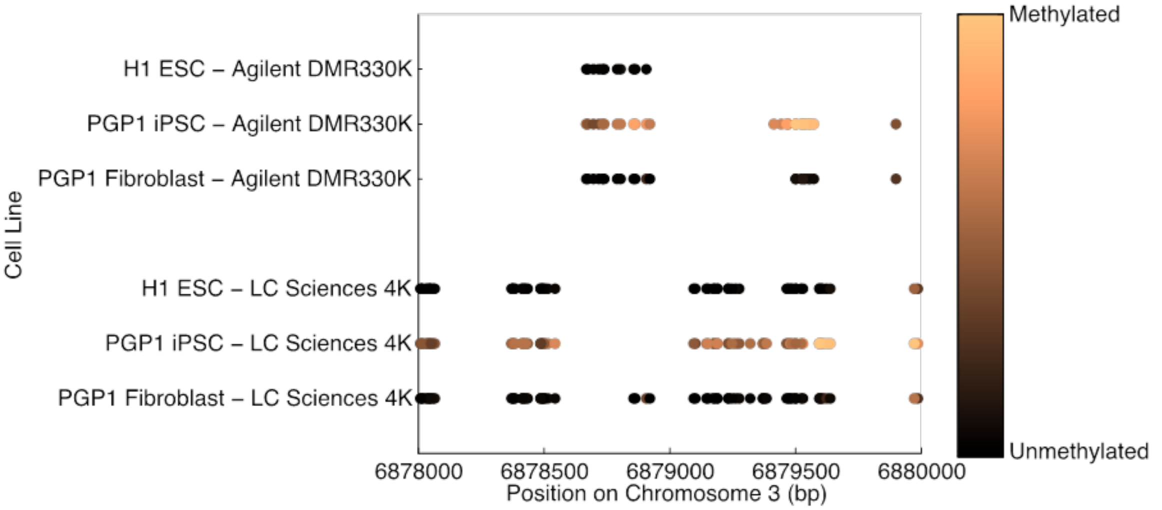


Figure 2.12 UCSC Genome Browser view showing an example of aberrant iPSC-specific methylation after reprogramming of PGP1 fibroblasts into iPS cells. Circles represent a location with measurable methylation state, with black indicating unmethylated and gold indicating methylated. The Agilent 330K probe set identified a small intronic region containing aberrant methylation in the iPS cells that are not present in either the fibroblast progenitors or a control hESC line. The LC Sciences 4K probe set was designed to characterize the methylation state upstream and downstream of this region. This focused assay revealed that the abnormal methylation also extended into the exonic region of GRM7.

Table 2.1 Comparison of bisulfite sequencing methods. The number of enzymatic reactions, number of purifications, cost per sample, and mapping rates for first-generation padlock probes, second-generation library-free padlock probes, reduced representation bisulfite sequencing (RRBS), and whole genome bisulfite sequencing (WGBS) are shown.

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

[1] Unlike other methods, in the library-free BSPP protocol size selection is typically performed on 48-96 pooled libraries.

[2] Includes the cost of ordering 400,000 synthesized probes from LC Sciences and reagents for preparing probes, bisulfite conversion, capture, and sequencing library preparation. Estimates assume that 10,000 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)

Table 2.2 Representative cost per sample for oligonucleotide synthesis, sequencing library construction, and Illumina sequencing.

Expected number of samples to be processed	Probe set sizes		
	4,000	**40,000**	**400,000**
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

Table 2.3 Primer sequences used for padlock probe production, padlock capture, sequencing library construction, and Illumina sequencing.

Primer name	**Primer sequences**
Primers used with Agilent Probes	
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’
SolSeq6.3.3 (Read1)	5’-TACACCACTCTCAGATGTTATCGAGGTCCGAC -3’
SolSeqV6.3.2r(Read2)	5’-GCTAGGAACGATGAGCCTCCAAC-3’
AmpR6.3IndSeq(IndexRead)	5’-GTTGGAGGCTCATCGTTCCTAGC-3’
Primers used with LC Sciences Probes	
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’

Table 2.3 Primer sequences used for padlock probe production, padlock capture, sequencing library construction, and Illumina sequencing, continued.

Primer name	**Primer sequences**
Switch.CA-2-RA.Ind76Sol	5'-CAAGCAGAAGACGGCATACGAGATAATAGGCGGTCTGCCATCGGGAAGCTGAAG-3'
CP-2-SeqRead1.x (Read1)	5'-TACACGCCTATCGGGAAGCTGAAG-3'
CP-2-IndSeq.x (IndexRead)	5'-ACACTACCGTCGGATGGCAGACCG-3'
CP-2-SeqRead1.y (Read1)	5'-TACACGCCTATCCGACGGTAGTGT-3'
CP-2-IndSeq.y (IndexRead)	5'-CTTCAGCTTCCCGATGGCAGACCG-3'

* Indicates a phosphorothioate bond

2.4 References

1. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y and Zhang K (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 27:353-60.

2. Liu GH, Barkho BZ, Ruiz S, Diep D, Qu J, Yang SL, Panopoulos AD, Suzuki K, Kurian L, Walsh C, Thompson J, Boue S, Fung HL, Sancho-Martinez I, Zhang K, Yates J, 3rd and Izpisua Belmonte JC (2011) Recapitulation of premature ageing with iPSCs from Hutchinson-Gilford progeria syndrome. Nature 472:221-5.

3. Liu GH, Suzuki K, Qu J, Sancho-Martinez I, Yi F, Li M, Kumar S, Nivet E, Kim J, Soligalla RD, Dubova I, Goebl A, Plongthongkum N, Fung HL, Zhang K, Loring JF, Laurent LC and Izpisua Belmonte JC (2011) Targeted gene correction of laminopathy-associated LMNA mutations in patient-specific iPSCs. Cell Stem Cell 8:688-94.

4. Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, Barbera AJ, Zheng L, Zhang H, Huang S, Min J, Nicholson T, Chen T, Xu G, Shi Y, Zhang K and Shi YG (2011) Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol Cell 42:451-64.

5. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA and Feinberg AP (2011) Increased methylation variation in epigenetic domains across cancer types. Nat Genet 43:768-75.

6. Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, Lee JH, Loh YH, Manos PD, Montserrat N, Panopoulos AD, Ruiz S, Wilbert ML, Yu J, Kirkness EF, Izpisua Belmonte JC, Rossi DJ, Thomson JA, Eggan K, Daley GQ, Goldstein LS and Zhang K (2011) Somatic coding mutations in human induced pluripotent stem cells. Nature 471:63-7.

7. Wang H, Chattopadhyay A, Li Z, Daines B, Li Y, Gao C, Gibbs R, Zhang K and Chen R (2010) Rapid identification of heterozygous mutations in Drosophila melanogaster using genomic capture sequencing. Genome Research 20:981-988.

8. Turner EH, Lee C, Ng SB, Nickerson DA and Shendure J (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 6:315-6.

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

10. Irizarry RA, Ladd-Acosta C, Wen B, Wu ZJ, Montano C, Onyango P, Cui HM, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S and Feinberg AP (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics 41:178-186.

11. Doi A, Park IH, Wen B, Murakami P, Aryee MJ, Irizarry R, Herb B, Ladd-Acosta C, Rho JS, Loewer S, Miller J, Schlaeger T, Daley GQ and Feinberg AP (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nature Genetics 41:1350-U123.

12. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B and Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315-322.

13. Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, Schifano E, Booth J, van Putten W, Skrabanek L, Campagne F, Mazumdar M, Greally JM, Valk PJ, Lowenberg B, Delwel R and Melnick A (2010) DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer Cell 17:13-27.

14. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM and Bejerano G (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28:495-501.

15. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM and Ecker JR (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471:68-73.

CHAPTER 3 QUANTITATIVE COMPARISON OF DNA METHYLATION BIOMARKER ASSAY

3.1 Abstract

I have participated in the BLUEPRINT Biomarker Benchmark project with the major goal of comparing techniques for quantification of DNA methylation biomarkers at specific loci using the same reference samples set. The performances of the assays were evaluated based on accuracy, specificity, flexibility, robustness, and cost structure. I have performed targeted bisulfite sequencing using bisulfite padlock probe (BSPP) to capture 1,072 assigned CpG sites in 32 samples with various characteristics. The raw data were reported and further analyzed by the bioinformatics team of the project for comparison with other quantitative assays performed by different expert groups. The assays performed included bisulfite pyrosequencing, Epityper, MethylLight, RainDrop bisulfite sequencing, and Infinium 450k bead array. The preliminary report from the comparison showed that bisulfite pyrosequencing and amplicon bisulfite sequencing assays had the highest correlation with the consensus measurements, and BSPP method had a good performance on average with many correlation coefficients, r, greater than 0.9 when compared with other assays, including the bisulfite pyrosequencing and amplicon bisulfite sequencing assays. In addition, BSPP method showed a higher throughput compared to other assays for assaying hundreds to thousands of genomic regions. This suggested that BSPP method has high potential for use in clinical diagnostics although the power, cost, and workflow remain to be evaluated.

3.2 Introduction

The potential of DNA methylation as a clinical biomarker has become increasingly accepted in recent years [1]. DNA methylation biomarker detection is promising in personalized medicine via molecular diagnostics and prognostics. To detect DNA methylation markers related to pathological status at specific loci, a broad range of assays has been developed. To be adopted to routine clinical diagnostics, an assay needs to be more accurate, sensitive, affordable, and robust. However, no systematic technology comparison has yet been performed that included more than 2-3 different assays or accounted for the importance of inter-laboratory robustness. Compared to gene expression signatures, which are already been used in routine diagnostic in the clinic, DNA methylation is more promising to be implemented as a part of clinical workflows. DNA is more stable than RNA, therefore detection DNA methylation in various types of samples such as blood plasma, FFPE material, or other body fluid is possible. The available techniques that have been widely used to determine DNA methylation at genome scale such as whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), or array-based assays, tend to be costly, labor-intensive, and impractical in the clinic. The goal of the BLUEPRINT Biomarker Benchmark project is to systematically compare established and emerging assays for locus-specific DNA methylation quantification in terms of their accuracy, sensitivity and specificity; their inter-laboratory consistency and robustness; their flexibility and ease of use; and their sample throughput and cost structure. To that end, a set of 32 reference samples has been established and analyzed for candidate biomarker loci with various characteristics. Aliquots of these reference samples were provided to experts in the field of DNA methylation biomarker development from around the world, who then measured locus-specific DNA methylation levels in these samples using their favorite assay(s). All data were cross-compared and analyzed against a standard reference by the project bioinformatics team, in order to derive assay-specific performance profiles. Partial of detailed results have been reported back to all participating labs, the remaining data analysis of some aspects are in progress, and the anonymized results will be presented in a joint publication.

In summary, this technology comparison will provide researchers, clinicians and regulatory agencies with quantitative data on the comparative performance of various DNA methylation assays and with confidence in developing and validating DNA methylation biomarkers for use in clinical diagnostics.

3.3 Materials and Methods

3.3.1 Padlock probes design

We modified our previous padlock probes design to contain a synthetic unique molecular identifier (UMI) tag consisting of ten random bases (Figure 4.1). We used ppDesigner to obtain bisulfite probes on both Watson and Crick strands for a target region up to 500 bp upstream and downstream of each CpG site. The algorithm applied by ppDesigner was used to generate padlock probes [2, 3]. The capture sequences were restricted to 25 to 30 bp, and not more than 54 bp in total length per probe. We selected one probe per strand that covers the target CpG sites with a fill-in size between 200-280 bp. We designed a total of 2,029 padlock probes with capture sequences containing between 0-6 CpG sites and an average of ~ 0.79 CpG per capture sequence. During probe assembly, we enumerated both methylated and unmethylated versions of the CpG sites within capture sequences, ending up with multiple probes per target. We obtained 4,400 uniquely assembled oligonucleotides that we then made 2-3 replicates of to fill 12,000 positions on a chip from CustomArray Inc.

3.3.2 Padlock probes library preparation

The oligonucleotides were prepared as described in 2.3.1.1 with some minor modifications. We have purchased synthetic oligonucleotides from a different vendor (CustomArray, Inc).

3.3.3 Sample preparation and BSPP capture

I performed bisulfite conversion and BSPP capture as described in 2.3.2 with minor modifications. I carried out bisulfite conversion on 500ng each of DNA sample using EZ-96 DNA Methylation Lightning MagPrep (Zymoresearch) following the protocol from manufacturer. We then measured bisulfite converted DNA with Qubit ssDNA assay, and also calculated recovery rate. Approximately 150ng of bisulfite treated DNA was mixed with BSPP in 1X AmpLigase buffer (Epicentre) in total volume 20µl . I overlaid the mixture reaction with 40µl mineral oil to prevent evaporation during padlock probe capture. The reaction was incubated on the thermocycler with the following thermocyling program; 95˚C for 30s, cool down to 54˚C at 0.02˚C/s, hold at 55˚C for 20hr. We then added 2µl of KLN solution mix to the capture reaction (KLN: 20% v/v Hemo KlenTaq (NEB), 0.5U/µl AmpLigase (Epicentre); 100uM dNTP, 1X AmpLigase buffer) and continued to incubate at 55˚C for 20hr to fill the gap between the capturing arms and to ligate resulting in circularized DNA. I denatured the DNA by heating at 94˚C for 2 min. I digested bisulfite converted DNA and free BSPP in the mixture by adding

1µl each of Exonuclease I (Epicentre) and Exonuclease III (Epicentre) and incubated at 37˚C for 2hr. We inactivated exonuclease enzymes by incubation at 94˚C for 5min and stored captured DNA at 4˚C or performed amplification immediately.

3.3.4 Amplification and sequencing library construction

I firstly amplified captured DNA in a small volume of 25µl to monitor the number of cycle to amplify and to verify if the capture work. 2.5µl of circularized DNA was added to 1X KAPA SYBR Fast qPCR Master Mix with 200nM each of AmpF6.4.Sol and AmpR6.3.Index primers in total volume 25µl and incubated the reaction at 98°C for 30s, 8 cycles of 98°C for 10s, 58°C for 20, 72°C for 20s, 15 cycles of 98°C for 10s, 72°C for 20, and 72°C for 3min. I verified 3µl of PCR product in 6% TBE gel. Once I got the optimal cycle number and obtained the right expected amplicon size between (375bp-480bp), I continued to perform PCR in a larger volume of 100 µl. I amplified the rest of captured DNA by adding 10µl of circularized DNA in total 50µl reaction with 200nM each of AmpF6.4.Sol and AmpR6.3.Index primer, 1X KAPA SYBR Fast qPCR Master Mix in duplicates, and incubated on thermocycler as the thermocycling program above. I pooled 100µl of PCR product, purified with 0.8 volume of AMPure bead, eluted with 60µl EB buffer, and verify 3µl in 6% TBE gel. I determined concentration of each library by PAGE quantification. I combined each library in the same pool with equimolar ratio and performed PAGE-size selection by cutting the smear between 475bp -500bp. I resuspended sequencing libraries with approximately 20µl nuclease-free water and performed qPCR to quantified concentration of the pooled libraries. I ran our libraries in Illumina MiSeq run (PE, 250bp + 6bp + 250bp) with SolSeq6.3.3 (Read1), SolSeqV6.3.2r (Read2), and AmpR6.3IndSeq (IndexRead) primers.

3.3.5 Bioinformatics analysis

We first extracted the UMI from the first 10 bases of read 1 to label both reads in the PE reads generated from sequencing. Then read 1 and read 2 were analyzed separately and in parallel. First, we trimmed the ends of the reads to remove adaptor sequences using fastq-mcf [4]. Second, the trimmed reads were analyzed using the BisReadMapper pipeline to align the reads to the genome with BWA [5]**.** Next, an in-house script was used to apply clonal removal using the UMI tags. Samtools version 1.18 was used to generate bam and pileup files. We used a second in-house script to extract methylation values at CpG sites from the pileup and generating the methylation frequency files. Finally, we merge the methylation frequency files for read 1 and read 2 together. However, for each CpG site on each strand, we kept only the data from the read with the higher coverage or read 1 if the coverages are equal. The data were reported to the BLUEPRINT project team for analysis to compare the performance of the assays.

3.4 Results

To verify the performance of BSPP method for measuring DNA methylation biomarker, we have designed the new set of bisulfite padlock probe (BSPP) containing 4,400 unique probes to cover CpG sites on both strands of gDNA in 16 mandatory loci, 32 recommended loci and 1,024 optional loci. Those targeted loci were on promotor CpG islands, DNase hypersensitive sites, gene regulatory elements, exonic regions, intronic region, intergenic regions, and repeats. I have performed library-free BSPP capture on 32 DNA samples (6 of human colon tumor, 6 of human normal colon tissue, 4 of human leukemia cell line, 6 of artificially methylated human DNA, 6 of pooled human cancer and normal blood cell, and 4 of human colon cancer cell line xenograph.) All 32 samples were included to assess for accuracy, sensitivity, specificity and robustness of the assay. I have performed the first experiment on control DNA and used the sequencing results to inform me of the efficiencies of each probe. There were 566 probes that had zero capture product and others with very low capture products. I have re-synthesized the 566 probes with zero capture product and 823 probes with 1-91 capture product and pooled the two subsets to the first probe pool by varying ratio of the probes in the pool to normalize the capture efficiency of the high efficiency and low efficiency probes.

I then have applied the normalized probe set on the 32 reference samples. I was able to capture 14 mandatory CpGs, 28 recommended CpGs, and 928 optional CpGs in average across all 32 samples, which accounts for 89% of total targeted CpGs (Table 3.1). This suggested the robustness of our assay to be able to capture almost of the assigned CpG targets. The quality of the data was determined by calculating the correlation of DNA methylation values between Watson and Crick strand. The majority of the samples showed highly correlation with Pearson's correlation coefficient R more than 0.9 (Table 3.1). In addition, the four samples of formalin fixed paraffin embedded (FFPE) tissue samples also showed a high correlation as well, which suggested that our assay is good enough to be applied for the difficult sample such as FFPE tissue samples. We noticed that the six samples contain *in vitro* methylated DNA spiked into unmethylated whole genome amplified (WGA) DNA (sample BP17-22, Table 3.1) had very low correlation with Pearson' correlation R less than 0.4. This was possible that probes predominantly captured an unmethylated WGA DNA that existed in a large fraction and caused poor correlation as all methylation values were zero or extremely low.

The assay comparison was performed by the bioinformatics team of the BLUEPRINT Biomarker Benchmark project and the preliminary results of all assays in terms of accuracy and consistency, between assays and between the labs performing the same assay, were shared to all participants. Bisulfite pyrosequencing method and amplicon bisulfite sequencing showed the best performance based on the similarity of DNA methylation level to the consensus measurement. However, there were variations between different groups, and the number of measured CpG sites was also varied for low (~50% of assigned targets) to high (>90% of assigned targets). It's possible that the high correlation observed in bisulfite pyrosequencing and amplicon bisulfite sequencing assays was the consequence of overabundance of these types of assays in the study. BSPP method was among the good assays and showed a high correlation to bisulfite pyrosequencing and amplicon bisulfite sequencing assays with Pearson's correlation coefficient greater than 0.9 (Figure 3.2). The noticeable performance of BSPP assay is the throughput as it can assess as high as 89% of the targeted CpGs in all samples (Table 3.1).

3.5 Conclusions

In conclusion, we have applied BSPP capture on a set of 32 DNA samples prepared from fresh frozen tissue and FFPE tissue of human normal tissue and tumor tissue. In addition, small amount of human cancer DNA were spiked in human normal blood cell DNA to test for the power of the assay for detecting DNA methylation signature of rare molecules in the DNA pool. The targeted CpG sites on different genomic locations with different characteristics were assigned. I have optimized the capture assay using two rounds of captures and sequencing to adjust the probe ratio for those probes with zero capture products and low efficiency. The performance profiles of all assays were compared based on the accuracy of the measurement to the consensus methylation values, and the measurement consistency between the assays or between the labs. In average, bisulfite pyrosequencing assay showed the best performance based on agreement with the consensus, although this may be due to the over-representation of these types of assays in the study. BSPP method was classified in a good assay group based on criteria above and showed a high correlation with other assays that agreed well with consensus. Meaning that a high level of accuracy can be achieved with BSPP and in a very high throughput manner. This study showed that targeted BSPP capture has a high potential as the assay to be used in the clinic for methylation marker detection based on accuracy and throughput.

Chapter 3, in full, is the study that was performed as a part of BLUEPRINT Biomarker Benchmark project. The current status of this project is in the progress of analyzing additional data to assess the power of each assay and to validate the workflow and cost structure. The consortium paper will be released after all validations are finished. I acknowledge Dinh H. Diep for contribution in this work in part of padlock probe design and data processing. I was a primary author and performed BSPP assay.

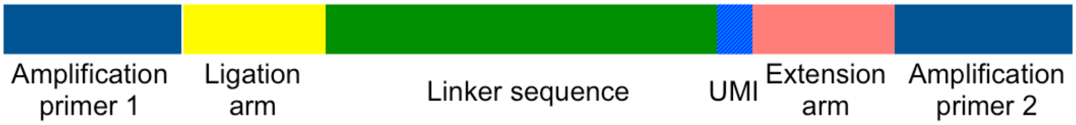
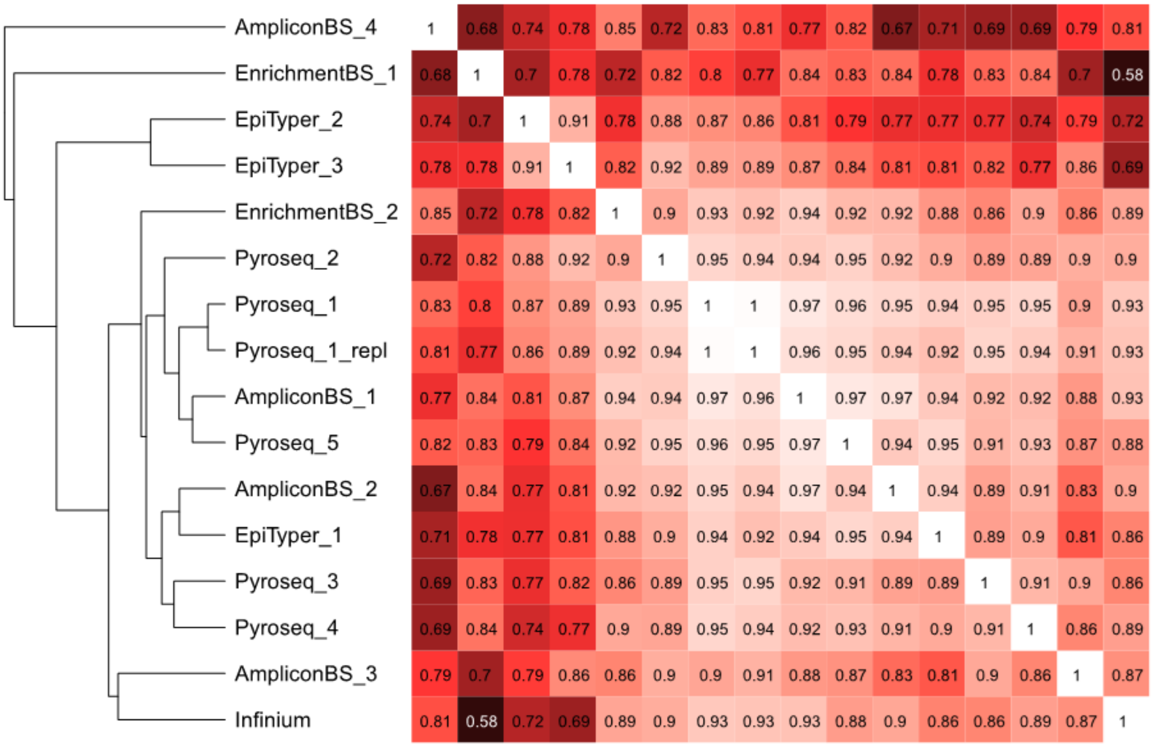


Figure 3.1 Schematic of padlock probe design



AmpliconBS_4	1	0.68	0.74	0.78	0.85	0.72	0.83	0.81	0.77	0.82	0.67	0.71	0.69	0.69	0.79	0.81
EnrichmentBS_1	0.68	1	0.7	0.78	0.72	0.82	0.8	0.77	0.84	0.83	0.84	0.78	0.83	0.84	0.7	0.58
EpiTyper_2	0.74	0.7	1	0.91	0.78	0.88	0.87	0.86	0.81	0.79	0.77	0.77	0.77	0.74	0.79	0.72
EpiTyper_3	0.78	0.78	0.91	1	0.82	0.92	0.89	0.89	0.87	0.84	0.81	0.81	0.82	0.77	0.86	0.69
EnrichmentBS_2	0.85	0.72	0.78	0.82	1	0.9	0.93	0.92	0.94	0.92	0.92	0.88	0.86	0.9	0.86	0.89
Pyroseq_2	0.72	0.82	0.88	0.92	0.9	1	0.95	0.94	0.94	0.95	0.92	0.9	0.89	0.89	0.9	0.9
Pyroseq_1	0.83	0.8	0.87	0.89	0.93	0.95	1	1	0.97	0.96	0.95	0.94	0.95	0.95	0.9	0.93
Pyroseq_1_repl	0.81	0.77	0.86	0.89	0.92	0.94	1	1	0.96	0.95	0.94	0.92	0.95	0.94	0.91	0.93
AmpliconBS_1	0.77	0.84	0.81	0.87	0.94	0.94	0.97	0.96	1	0.97	0.97	0.94	0.92	0.92	0.88	0.93
Pyroseq_5	0.82	0.83	0.79	0.84	0.92	0.95	0.96	0.95	0.97	1	0.94	0.95	0.91	0.93	0.87	0.88
AmpliconBS_2	0.67	0.84	0.77	0.81	0.92	0.92	0.95	0.94	0.97	0.94	1	0.94	0.89	0.91	0.83	0.9
EpiTyper_1	0.71	0.78	0.77	0.81	0.88	0.9	0.94	0.92	0.94	0.95	0.94	1	0.89	0.9	0.81	0.86
Pyroseq_3	0.69	0.83	0.77	0.82	0.86	0.89	0.95	0.95	0.92	0.91	0.89	0.89	1	0.91	0.9	0.86
Pyroseq_4	0.69	0.84	0.74	0.77	0.9	0.89	0.95	0.94	0.92	0.93	0.91	0.9	0.91	1	0.86	0.89
AmpliconBS_3	0.79	0.7	0.79	0.86	0.86	0.9	0.9	0.91	0.88	0.87	0.83	0.81	0.9	0.86	1	0.87
Infinium	0.81	0.58	0.72	0.69	0.89	0.9	0.93	0.93	0.93	0.88	0.9	0.86	0.86	0.89	0.87	1

Figure 3.2 Similarity and differences between assays. BSPP assay was assigned as EnrichmentBS_2.

Table 3.1 Watson and crick strand correlation at the same CpG site and number of captured CpGs.

SampleID	**fr_N**	**Correlation**	**#Mandatory(1x)**	**#Recommended (1x)**	**#Optional (1x)**	**Fraction target**
BP1	2,321	0.885	14	27	906	86%
BP2	2,433	0.942	14	29	916	88%
BP3	2,128	0.910	14	28	904	86%
BP4	1,344	0.944	14	25	901	86%
BP5	2,933	0.919	12	28	943	90%
BP6	2,595	0.949	14	29	933	89%
BP7	2,874	0.908	15	29	945	90%
BP8	3,145	0.945	12	29	928	88%
BP9	2,832	0.875	14	30	948	91%
BP10	2,438	0.931	13	27	915	87%
BP11	2,209	0.909	14	28	940	90%
BP12	1,859	0.939	13	27	921	88%
BP13	3,226	0.959	14	29	937	89%
BP14	2,800	0.905	14	28	935	89%
BP15	1,621	0.967	12	25	896	85%
BP16	2,009	0.824	13	27	926	88%
BP17	3,502	0.294	14	28	932	89%
BP18	3,625	0.323	15	29	945	90%
BP19	2,843	0.187	14	27	913	87%
BP20	3,718	0.389	15	29	949	91%
BP21	3,477	0.309	14	28	954	91%
BP22	3,176	0.128	15	30	949	91%
BP23	3,333	0.941	15	29	941	90%
BP24	2,930	0.904	15	29	935	89%
BP25	2,623	0.959	13	28	935	89%
BP26	2,475	0.961	15	29	939	90%
BP27	2,592	0.966	14	29	925	88%
BP28	2,133	0.953	13	26	918	87%
BP29	2,400	0.944	15	30	936	90%
BP30	3,429	0.960	14	30	954	91%
BP31	814	0.917	14	28	872	83%
BP32	1,234	0.932	14	30	889	85%
Average		**0.806**	**14**	**28**	**928**	**89%**

3.6 References

1. Laird PW (2003) The power and the promise of DNA methylation markers. Nat Rev Cancer 3:253-66.

2. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y and Zhang K (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 27:353-60.

3. Diep D, Plongthongkum N, Gore A, Fung HL, Shoemaker R and Zhang K (2012) Library-free methylation sequencing with bisulfite padlock probes. Nat Methods 9:270-2.

4. Aronesty E and ea-utils (2011) Command-line tools for processing biological sequencing data.

5. Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589-95.

CHAPTER 4 CHARACTERIZATION OF GENOME-METHYLOME INTERACTIONS IN 22 NUCLEAR PEDIGREES

4.1 Abstract

Genetic polymorphisms can shape the global landscape of DNA methylation, by either changing substrates for DNA methyltransferases or altering the DNA binding affinity of *cis*-regulatory proteins. The interactions between CpG methylation and genetic polymorphisms have been previously investigated by methylation quantitative trait loci (mQTL) and allele-specific methylation (ASM) analysis. However, it remains unclear whether these approaches can effectively and comprehensively identify all genetic variants that contribute to the inter-individual variation of DNA methylation levels. Here we used three independent approaches to systematically investigate the influence of genetic polymorphisms on variability in DNA methylation by characterizing the methylation state of 96 whole blood samples in 52 parent-child trios from 22 nuclear pedigrees. We performed targeted bisulfite sequencing with padlock probes to quantify the absolute DNA methylation levels at a set of 411,800 CpG sites in the human genome. With mid-parent offspring (MPO) analysis, we identified 10,593 CpG sites that exhibited heritable methylation patterns, among which 70.1% were SNPs directly present in methylated CpG dinucleotides. We determined the mQTL analysis identified 49.9% of heritable CpG sites for which regulation occurred in a distal *cis*-regulatory manner, and that ASM analysis was only able to identify 5%. Finally, we identified hundreds of clusters in the human genome for which the degree of variation of CpG methylation, as opposed to whether or not CpG sites were methylated, was associated with genetic polymorphisms, supporting a recent hypothesis on the genetic influence of phenotypic

plasticity. These results show that *cis*-regulatory SNPs identified by mQTL do not comprise the full extent of heritable CpG methylation, and that ASM appears overall unreliable. Overall, the extent of genome-methylome interactions is well beyond what is detectible with the commonly used mQTL and ASM approaches, and is likely to include effects on plasticity.

4.2 Introduction

DNA methylation represents an important layer of epigenetic regulation on the transcriptional activity of the human genome and plays a crucial role in genomic imprinting, embryonic development and determination of cell type. Accumulating evidence suggests that DNA methylation patterns, rather than being similar within members of the same species, vary from one individual to another [1-3] due to both genetic and environmental factors [4, 5]. This variability could potentially explain why certain phenotypic outcomes manifest differently across individuals of the same species, including in terms of the susceptibility to and treatability of many human diseases [6, 7].

With the recent advances in DNA methylation assays, a growing number of studies have identified a genetic contribution to inter-individual variation in DNA methylomes. One type of study relies on methylation quantitative trait locus (mQTL) mapping, which identifies genomic polymorphisms associated with variation of CpG methylation in a *cis*-regulatory manner [8-11]. An alternative approach involves characterizing allele-specific methylation, in which a change in a specific polymorphism leads to the direct loss or gain of DNA methylation [2, 3, 12-15]. While an increasingly large number of associations between SNPs and CpG sites have been reported in these recent efforts, it remains unclear whether mQTL and ASM analyses are truly uncovering the full extent of genome-methylome interactions. In this study, we performed targeted bisulfite sequencing on human whole blood samples from 96 individuals representing 22 nuclear pedigrees, and took advantage of the parent-child trios using mid-parent offspring (MPO) analysis to fully uncover genome-methylome interactions. We then performed mQTL and ASM analysis on the same samples, and investigated the capability of each method to identify the genetic contribution to inter-sample methylation variability.

4.3 Materials and Methods

4.3.1 Sample collection

Genomic DNAs from the 96 individuals of 22 pedigrees were extracted from whole blood previously collected as part of an on-going genetic study of schizophrenia under the IRB approvals by Utrecht and UCLA. Written consents were obtained from all donors. All personal identifiers were removed and replaced by alpha numerical codes for sample tracking. The information that is available to us as researchers include age, gender and family relationships.

4.3.2 Targeted bisulfite sequencing with padlock probes

Bisulfite padlock probe design, production and sequencing were previously described[16, 17]. Briefly, genomic DNA was extracted from peripheral blood of 22 pedigrees, and approximately 1 µg of genomic DNA was bisulfite converted with EZ-96 Zymo DNA Methylation-Gold kit (Zymo Research). Approximately 250ng of bisulfite converted genomic DNAs were mixed with normalized amount of genome-wide scale padlock probes and oligo suppressors. The padlock probes were annealed to bisulfite converted genomic DNA. The gap between two ends of padlock probes was filled and ligated with AmpliTaq DNA polymerase, Stoffel fragment (Life Technologies) and Ampligase (Epicentre), respectively resulting in circularized DNA. The bisulfite sequencing libraries were generated by library-free BSPP protocol as described[17]. Briefly, two-thirds of the circularized DNA of each captured reaction were directly amplified and barcoded with adapter primers compatible with Illumina sequencer. The bisulfite sequencing libraries were purified with AMPure XP magnetic beads (Agencourt), pooled in equimolar ratios, size selected at the size approximately 375bp with 6% TBE polyacrylamide gel (Life Technologies), and sequenced by Illumina HiSeq2000 and GAIIx sequencers.

4.3.3 DNA methylation data

The pooled libraries were firstly sequenced with Illumina HiSeq2000 sequencer (100bp, paired-end reads). Additional sequencings were performed for those samples with number of reads less than 22 millions (53 samples) on the same sequencing libraries with Illumina HiSeq2000 and GAIIx sequencers. Bisulfite sequencing data were processed as described[13, 17]. Briefly, adapter sequences (27bp from 5' end) were trimmed from bisulfite reads prior to mapping. In bisulfite sequencing reads, all cytosines were replaced by thymines and mapped to the *in silico* bisulfite converted human genome sequences (hg19) with all cytosines converted to thymines on both strands by bisReadMapper[17]. Absolute DNA methylation level at each CpG site with minimum 10X depth coverage in each sample was calculated at level from 0-1. Summary statistics for sequencing read mapping for all samples sample were reported in Supplementary file 7. The quality of the data was assessed by comparing DNA methylation levels at the same CpG sites captured and measured independently on the two strands, which can be treated as internal technical replicates.

4.3.4 Mid-parent offspring analysis

Mid-parent offspring (MPO) analysis was performed by mid-parent offspring regression[18] to estimate the heritability of DNA methylation at each CpG site. DNA methylation level of the offspring in each trio was compared against the mean DNA methylation level of the parents. In total, 76,408 autosomal variable CpGs (minimum standard deviation of 0.1) shared in at least 80% of subjects were analyzed. The slope of the fitted line was used to estimate the heritability (h^2) of each CpG site. CpG sites with h^2 greater than 0.2 in a minimum sample size (number of trio) of 10 were defined as heritable CpGs. The Benjamini-Hochberg method was used to correct for multiple testing errors.

4.3.5 Methylation quantitative trait loci

Methylation quantitative trait loci (mQTL) analysis was performed by PLINK[19] to determine the association between DNA methylation level of variable CpG sites as described above and SNP genotypes called from methylation data (15,450 SNPs) of 96 subjects or imputed autosomal SNP genotypes (5,257,772 SNPs) of 57 subjects generated by Illumina SNP array (550K) and Affymetrix SNP array. SNP genotypes with a minor allele frequency (MAF) of at least 0.05 and with a Hardy-Weinberg Equilibrium (HWE) p-value > 0.001 were included in this analysis. Mendel error rates in each nuclear family with the full trio were calculated by PLINK (Table 4.S6) We used least square linear regression, and the corresponding p-values were calculated for each CpG-SNP association pair within 1Mb. FDR was calculated by Benjamini-Hochberg multiple correction method to assess the significance of the CpG-SNP association. To deal with family structure, QFAM analysis was performed. 10,000 permutations were performed and p-value was empirically calculated as the fraction of permuted data test-statistic is larger than the non-permuted data test statistic. Additional analyses were performed on subsets of imputed SNPs including 618,580 index SNPs present on Illumina 1M SNP array. The SNPs that showed strong correlation with DNA methylation were extracted and annotated significant QTL as *cis* if the SNP lay within 1 Mbs of the CpG site.

4.3.6 SNP imputation

Array genotype data of 96 subjects of this study were generated on two different array platforms, 23 individuals on Illumina SNP array (550K) and 73 individuals on Affymetrix SNP array by Wellcome Trust Case Control Consortium 2 (WTCCC2). After removing poor quality genotyping, there were SNP data of 57 subjects in this study (11 individuals on Illumina SNP array and 46 individual on Affymetrix SNP array). There were 150K of SNP overlapping between the two platforms, so imputation was performed on the two data sets independently. For Illumina SNP data, SNP genotype data from unrelated individuals were phased with Beagle[20] then imputed with Minimac[21] with the 1000 Genomes Project reference[22]. After post-imputation quality control, there were total imputed 8,064,119 SNPs (MAF of 0.01, r^2 of 0.3). For Affymetrix data set, the SNP genotypes of 43 individuals were imputed with SNP data genotyped on Affymetrix SNP array, including 268 pairs, 236 trios, and 926 unrelated individuals. All Mendel inconsistencies were set to missing before phased with Beagle to take into account family structure. Then Minimac was used for imputation. There were 8,022,142 SNPs after the post-imputation quality control. Approximately 7,800,000 overlapping SNPs between the two imputed data sets were merged by including only well imputed SNPs on the two data sets. SNPs with MAF > 0.05 and HWE >0.001 were extracted, and there were 5,257,772 imputed SNPs remained in this study.

4.3.7 Allele-specific methylation

Allele-specific methylation (ASM) analysis was performed as described[13]. Briefly, we generated the 2 X 2 contingency table where the two columns containing the two alleles and the two rows containing the counts of methylated and un-methylated cytosines at CpG site(s) on the read containing heterozygous SNP(s). The p-value at each CpG site was calculated by Fisher's exact test. We identified ASM if the p-value was less than 0.001 and the methylation frequency between the two alleles was greater than 0.2.

4.3.8 Genomic region annotation

Genomic features of CpG sites were assigned using bedtools[23] according to genomic annotation structure described by Bikikova et al, 2011[24]. The enrichment of CpG sites from different analyses was calculated as the ratio between significant CpG sites from each analysis and CpG sites included in the analysis.

4.3.9 Variation-SNP and variably mathylated regions

We identified vSNPs and VMRs by performing association tests. Linear regression was performed on the variance of DNA methylation at each CpG site among individuals and the three genotype groups (AA, AB, BB) within 1Mb distance. The t-score of each CpG-SNP pair was calculated, and the false discovery rate was calculated by using different cutoff values for the test statistic values. To deal with the high rate of false positive signals, we required at least five adjacent CpG sites with maximal spacing 200 bp between CpGs showing consistent association for VMRs. We then grouped the overlapping or adjacent VMRs into clusters. We note that VMRs associated with different vSNPs could be partially overlapping, so they could be grouped into the same cluster.

4.4 Results

We characterized DNA methylation levels in genomic DNA from the peripheral blood of 96 individuals in 22 nuclear pedigrees of European ancestry, each including one proband with schizophrenia, two unaffected parents and one or two unaffected siblings (a total of 52 trios of two parents and one child). We measured CpG methylation at single base resolution using ~330,000 bisulfite padlock probes capturing a pre-selected subset of genomic regions, including promoters, enhancers, DNase I hypersensitive sites and other regions known to be variable among different cell types [17]. Note that, like other bisulfite-based methods, 5-methylcytosine and 5-hydroxymethylcytosine are indistinguishable with this assay. In addition, several recent works have shown that variation in cell composition is a confounding factor[25-27]. In this study, we did not correct for cell composition due to the lack of reference data from pure cell populations, and treated the average methylation of all cells in whole blood as a quantitative trait. On average, we obtained methylation measurements for ~500,000 CpG sites per sample. A total of 411,800 autosomal CpG sites (and 5,133 on sex chromosomes) had valid methylation measurements in at least 80% of samples. We filtered out CpG sites showing low variability among samples ("static CpG sites"), and focused all further analysis on a subset of 76,408 autosomal variable CpG sites (those with standard deviation of methylation levels across all samples ≥ 0.1). Hierarchical clustering based on the methylation levels of highly variable autosomal CpG sites (standard deviation ≥ 0.3) showed a clustering pattern consistent with the family structure (Figure 4.S1). While several samples came from individuals with schizophrenia, the sample size here was too small to perform any significant association tests between disease state and either genetic or methylation factors; thus, we focused on treating methylation itself as a quantitative trait and investigating its relation to individual genetic variants.

MPO identifies CpG sites known to have heritable methylation patterns using trio information

In order to obtain an independent list of CpG sites where variability in DNA methylation was known to be related to genetic factors, we performed mid-parent offspring (MPO) analysis [18], which analyzes the correlation between the mean methylation level at each CpG site in each parent pair and the methylation level at the same CpG sites in the child (Figure 4.1a). This family-based analysis of each trio allowed identification of any potential heritable methylation patterns irrespective of the type and frequency of genetic variants (i.e. SNPs, indels, structural genomic variation) or the method of regulation. We identified CpG sites as heritable by requiring a heritability (h^2) value greater than 0.2 in a minimum of available data in ten trios with a FDR cutoff of 0.05 (with Benjamini-Hochberg correction).

We identified a total of 10,593 CpG sites that possessed variable methylation directly correlated with genetic pedigree (Supplementary file 1), accounting for ~13.9% of all variable CpG sites. This result suggests, based on the samples in this study, that genetic factors account for over ten percent of inter-sample DNA methylation variability in human blood. Further analysis revealed that 70% (7,424) of these CpG sites in fact showed variable methylation due to their containing a family-specific SNP at exactly the same locus. This result indicates that the majority of heritable CpG methylation patterns are due to genetic polymorphisms directly altering the substrates of DNA methyltransferases (“SNP-CpGs”), whereas other *cis*- or *trans*- regulatory effects account for only a small fraction (3,169, ~30%) of heritable CpG methylation ("non-SNP CpGs") (Figure 4.2a). Non-SNP CpG sites that localized close by appeared to share similar methylation patterns within individuals of the same family, suggesting that one genetic variant or haplotype could be affecting multiple CpG sites (Supplementary file 2, Figure 4.1b-c). Heritable CpG sites were not enriched for any particular genomic region, as they showed a similar distribution across the genome as all variable CpG sites (Table 4.S1). However, moderate enrichment in gene body and intergenic regions was observed over all characterized CpGs. (Table 4.S1)

mQTL finds associations between SNPs and CpG sites in a population without trio information

While it is possible to identify heritability in DNA methylation through MPO analysis, for a majority of cases, parent-child trio data is unavailable. In order to determine what fraction of genome-methylome interactions could be identified at a population level when pedigree information was not present, we treated each CpG site as a methylation quantitative trait locus (mQTL), and analyzed the effects on methylation levels of common SNPs or other genetic variants in linkage disequilibrium (LD) with the index SNPs. We sought to perform an analysis using SNP genotypes determined by multiple platforms in order to identify the optimal strategy for identifying genomic contributions to methylation. In some cases, performing additional experiments to obtain sample genotypes is cost-prohibitive; we therefore first utilized the bisulfite sequencing data itself to call genomic SNPs using a previously described method[17]. We obtained genotypes at 15,450 SNP sites after requiring genotypes to be called at putative SNP sites in at least 75% of subjects. Because these SNPs were called only in the captured regions, SNP density was low compared to the whole genome. In order to also perform a more comprehensive mQTL mapping using additional SNPs, we derived SNPs of 57 subjects, a subset of the 96 samples passing quality control of SNP genotyping, using both Affymetrix and Illumina SNP arrays. To avoid platform-specific technical differences, we performed imputation using SNP data from the 1,000 Genomes Project[22], and obtained genotypes for ~5 million SNPs per sample.

We performed mQTL regression analysis using PLINK with QFAM familial dependence correction [19] between the DNA methylation level of each variable CpG site and the genotypes of SNPs located up to 1 Mb upstream and downstream. Using SNP calls from the bisulfite sequencing data, we identified 7,593 CpG-SNP *cis*-associations at <5% FDR (Supplementary file 3), consisting of 4,253 CpG sites associated with 3,842 SNPs. With the ~5 million genome-wide SNPs, we identified a total of 644,773 CpG-SNP *cis*-associations at <5% FDR (Supplementary file 4), consisting of 9,783 CpGs associated with 412,382 SNPs. As in the MPO analysis, a majority of CpG-SNP interactions were due to genetic mutations directly at the CpG site (66.7% and 70.5%, respectively, Figure 4.2b, 4.2c).

Generally, the majority of *cis*-regulatory SNPs were located very close to their associated CpG sites in both SNP data sets. For the SNPs called from bisulfite sequencing reads, 47.6% of the CpG-SNP associations were within 2kb (Table 4.S2, Figure 4.S2a), and only 15.2% of associations were further away than 100kb (Table 4.S2, Figure 4.S2b, 4.S2e). For the SNPs called using genome-wide arrays that more uniformly capture the LD blocks in the human genome, over 64.9% of CpG-SNP associations were within 100kb (Table 4.S3, Figure 4.S2f), with the strongest associations mostly within 2kb (Table 4.S3, Figure 4.S2c). The identified additional enrichment of short-range CpG-SNP associations in the bisulfite sequencing SNP data appeared to be partially due to sampling bias, because SNPs were called only in captured regions and thus tended to locate very close to CpG sites (Figure 4.S2a, 4.S2e); it appears that to fully characterize long-range CpG-SNP interactions, SNP genotyping is required. However, bisREAD SNPs can be called directly from methylation sequencing data, whereas SNP genotyping experiments involve extra experimental cost. Additionally, even though the number of bisREAD SNPs used in our analysis was ~340 fold less than the genome-wide SNPs, it was still possible to identify half of the long-distance non-SNP CpG interactions. Therefore, in cases where SNP genotyping experiments are difficult to perform due to either limited biological material or budgetary constraints, SNPs called from bisulfite sequencing data can still be used to capture a reasonable fraction of *cis*-regulatory interactions, with the caveat that long distance interactions will be under-represented.

Finally, in order to ensure that CpG-SNP interactions were not being missed due to excessive penalties from multiple testing correction in the 5 million SNP case, we additionally performed mQTL analysis using a subset containing 618,580 SNPs in unique LD blocks. The number of CpG-SNP associations decreased to 67,781 (at FDR <5%), indicating that multiple testing penalties were not having a large impact on statistical testing in this case (as a similar fraction of CpG-SNP interactions out of total putative interactions were identified as true in each case).

ASM finds associations between SNPs and CpGs in single samples

We next used a third strategy to examine the attempt to discern the influence of genetic variation on DNA methylation levels by analyzing allele-specific methylation (ASM). Unlike the MPO and mQTL analysis methods, which utilize information from multiple samples together, ASM examines genome-methylome interactions in one sample at a time. Using this recently developed computational procedure [13], we identified an average of 2,266 variable CpG sites per individual that exhibited significant difference in allelic methylation based on genomic factors (methylation difference >0.2). Consistent with previous observations [12, 13, 28], most ASM events were due to SNPs present directly at CpG sites, (69.7%-92.5%, average 86.4%), with non-SNP CpG sites representing a very small fraction of putative genome-methylome interaction (Figure 4.S3a, 4.S3b). Additionally, the majority of detected ASM events were present in only a small fraction of subjects (Table 4.S4). After combining all overlapping ASM events, we identified 10,927 and 14,809 ASM events at non-SNP CpGs and SNP-CpGs respectively (Figure 4.2d). We observed a modest enrichment of ASM on non-SNP CpGs in gene body and intergenic regions (Table 4.S5, Figure 4.S3c, 4.S3d).

The efficacy of mQTL and ASM in identifying genome-methylome interaction

While the genomic *cis*-regulated CpG sites identified by MPO appear to be truly heritable through the use of trio information, it remained unclear to what extent mQTL and ASM analyses were characterizing true genome-methylome interactions. We thus next compared the three analyses to determine the efficacy of mQTL and ASM analysis.

While, as expected, most SNP-CpG sites identified by mQTL were true positive sites showing heritable CpG methylation (85.3%, Figure 4.S4a), surprisingly, only 49.9%

of non-SNP CpGs identified by mQTL analysis were found heritable by MPO analysis (Figure 4.3a), indicating that only half of non-SNP CpG sites identified by mQTL mapping are truly heritable. mQTL also failed to identify 54.6% of true heritable non-SNP CpGs (Figure 4.3a), indicating that for non-SNP CpGs, in addition to having a high false positive rate, mQTL analysis also appears to have a high false negative rate as well. This discrepancy could be due to a number of reasons, including lack of statistical power due to limited sample size, presence of long-range *cis*-interactions at a distance of over 1 megabase and/or *trans*-interactions [29], and the effects of other common or rare alleles not in LD with the SNPs tested. In addition, some marginally significant sites might be included or excluded due to the specific choices of p-value cut-offs for each of the two methods. In fact, when we plotted the mQTL association signals for heritable and non-heritable CpG sites separately, the majority of CpGs most strongly associated with SNPs (low p-value) were heritable CpGs (Figure 4.3b, Figure 4.S4b). Non-heritable CpGs in general showed weaker association signals, especially for longer-range *cis*-interactions (Figure 4.3c, Figure 4.S4c). It is possible that heritable CpG sites not identified by mQTL analysis could be regulated by other genetic mechanisms.

In contrast to the mQTL analysis, only very small fractions of CpG sites that seemed to exhibit ASM in at least one sample were found to be heritable (5.6% for non-SNP CpGs, 32.6% for SNP-CpGs) (Table 4.S4). One possibility is that calls made by ASM contain a high number of false positive CpG-SNP interactions. However, when we restricted our analysis to the CpG sites that exhibited consistent ASM patterns in two or more individuals, the fractions of sites overlapping with heritable CpGs increased only moderately, and remained far from the 49.9% or 85.3% overlap observed between mQTL calls and heritable CpGs. These calls could be explained by a number of possibilities, including non-genetic parent-of-origin effects (including but not limited to imprinting), random allelic drift [30], environmental factors, potentially higher false positive rates, or higher sensitivity than MPO in detecting allelic differences. Overall, however, ASM appears to have very low specificity in identifying CpG sites regulated by genetic variants.

Genetic polymorphisms affect the degree of variability in DNA methylation

Recently, it was proposed that genetic variants might be regulating the level of variability in molecular phenotypes such as CpG methylation rather than just regulating the exact methylation state[31, 32]. Under this hypothesis, a particular allele of a SNP is associated with highly variable methylation patterns across multiple individuals (Figure 4.4b) as opposed to being associated with a consistent increase or decrease in mean methylation level (Figure 4.4a). To determine if variation-SNPs (vSNPs) were present in this data set, we performed a regression analysis on the variance of DNA methylation at each CpG site and the genotypes of nearby SNPs (within 1Mb). A major technical challenge is that there are only three genotypes for each SNP, and hence the sample size for each regression is limited to three; this could potentially result in a very high false positive rate. To counteract this, we required that a candidate vSNP had a consistent effect on at least five adjacent CpG sites. The false positive rate was estimated to be ~10% by applying the same procedure to randomly permuted methylation data.

A total of 1,058 genomically-linked variably methylated regions (VMRs) were identified, with many SNPs associated with the variance of multiple nearby CpG sites (Supplementary file 5, Figure 4.4a, 4.4b). These nearby sites were further grouped into

383 VMR clusters (Supplementary file 4.6) by combining multiple VMRs that were within 100kb. The majority of VMR clusters (316 clusters, 82.5%) were located within 1 Mb of a set of 438 genes. The largest VMR cluster involved 53 variable CpG sites in a 38kb region covering GNAS, which is a well documented imprinted gene that has a highly complex expression pattern from both strands[33, 34]. Two other large VMR clusters overlapped with the HoxA gene cluster and protocadherin gamma gene cluster, both of which contain multiple functionally related and co-regulated genes and pseudogenes.

While the full functional consequences of such variable methylation remain largely unknown, we note that very recently four SNPs were found to be associated with rheumatoid arthritis and variance of methylation [26]. In order to test whether the observed VMR clusters could translate into genotype-specific variation at the gene expression level, we examined the top 10 VMR clusters and their respective genes in an array-based whole blood gene expression data set of 240 independent subjects [35]. Nine of the genes within the top ten VMR clusters were expressed at detectable levels (Table 1). Even though the effect sizes were small, we observed three genes (*GNAS*, *PEG3*, and *PCDHGA5*) from different VMR clusters all showing genotype-specific differences contributing to variance at the gene expression level.

4.5 Discussion

In the recent years, association mapping of molecular phenotypes such as gene expression, DNA methylation, or chromatin accessibility as quantitative traits (eQTL, mQTL, dsQTL) has revealed how genetic variants contribute to inter-individual variability and provided additional insights into the modulation of disease susceptibility

[1, 18, 36-39]. The recent technical advances in low-cost genome-wide DNA methylation assays (such as the Illumina 450k methylation array [24], RRBS [40], and BSPP [17] have catalyzed a new wave of epigenome-wide association studies aiming to characterize the contribution of both genetic and environmental factors to disease susceptibility [4, 41], with encouraging progress already in sight [26, 42-44]. However, while new analysis techniques have connected genetic variants, CpG methylation, and disease phenotypes, it remains unclear to what extent we should expect interaction to occur between genetic variation and the variability of DNA methylation, what fraction of interactions are able to be captured with current approaches, and what strategy we should use to efficiently capture these interactions.

In this study, we revealed that a large extent of genome-methylome interaction is completely missed by current analysis methods. By comparing the results from mQTL analysis to MPO analysis, which is guaranteed to find heritable methylation patterns, in 22 nuclear pedigrees, we demonstrated that a large fraction of heritable traits affecting CpG methylation remain hard or impossible to detect with the most widely used analysis method. However, we hypothesize that *trans*-regulation might account for the majority of heritable CpG sites not detectible by conventional mQTL analysis. While the anti-correlation of promoter DNA methylation and gene expression has been observed for many years, the exact mechanistic explanation behind DNA methylation regulating gene expression has yet to be firmly established. More recent observations of positive correlation between gene-body methylation and gene expression have added additional confusion to the functional role of DNA methylation [16, 45-47]. Stadler et al. recently demonstrated that binding of protein factors to DNA can lead to local reduction of DNA methylation[48], providing the first direct evidence that DNA methylation in general is a passive mark for protein-DNA binding. A corollary of this observation is that a DNA binding protein (such as a transcription factor) for which the expression is an eQTL (i.e. regulated by a genetic variant) can affect DNA methylation levels in hundreds to thousands of its binding regions genome-wide. As such, a single functional variant might regulate many mQTLs, mostly in trans, mediated by its primary effect on a single transcription factor. Connecting these mQTLs to functional variants therefore cannot be accomplished by simple association tests using nearby CpGs and SNPs. Additional information on the transcriptional factors and their direct regulating genes would be required, such as that becoming increasingly available through large-scale ChIP-Seq and DHS mapping efforts like the ENCODE project [49]. A coherent statistical framework for association testing that incorporates the information of protein-DNA binding from genome-wide assays would also be necessary to fully explore genome-methylome interactions.

We also provided a practical assessment on the sensitivity of mQTL mapping at various SNP densities, showing that using over a large number of SNPs can improve the level of statistical significance with diminishing gains in detecting additional SNP-associated CpG sites. On the other hand, for projects based on bisulfite sequencing, the SNP genotypes called from the sequencing reads alone can be used to recover a reasonable fraction of associated CpG sites. As bisulfite sequencing is being widely adopted and algorithms for SNP calling from bisulfite data are being optimized [50], using the smaller number of obtained SNPs could represent an economical option for large-scale EWAS studies, with the understanding that a denser SNP map would still be necessary to recover the majority of long-range regulatory effects.

We additionally characterized the ability of ASM to identify heritable methylation patterns. While we found many CpG sites that both exhibited allele-specific methylation in different individuals and showed heritable methylation patterns across all the pedigrees, the majority of CpG sites identified in our ASM analysis could not be explained by consistent effects of *cis*-regulatory variants across multiple individuals. We reason that ASM analysis is more susceptible to many non-genetic factors, including parent-of-origin effects, random allelic drift, and technical artifacts, and hence might not be appropriate as a primary approach for identifying methylation traits regulated by genetic variants. Population level analysis such as mQTL or MPO (if trio information is available) appears to be necessary to accurately characterize genomic effects on methylation patterns.

Finally, we provide evidence supporting a recently proposed hypothesis that genetic variants can regulate not only the mean but also the variation of molecular phenotypes such as CpG methylation or gene expression. This is not unexpected, as gene regulatory networks are connected through both positive and negative feedback [51, 52]. Reduction of negative feedback has been shown to increase the variability in both prokaryotic and eukaryotic organisms [53, 54], lending mechanistic support to the idea that genetic variants affecting the strength of negative regulation could result in a difference in variability for the components involved in a molecular network. Feinberg and colleagues have proposed that epigenetic variability provides a mechanism for selectable phenotypic variation [32], and provided examples of variable DNA methylation and its role in cancer [31] and rheumatoid arthritis [26]. Although the full extent of variable DNA methylation, as well as its phenotypic consequences, remain to be further characterized with larger cohorts of genetically unrelated individuals, the observation of hundreds of VMRs in the 22 nuclear pedigrees analyzed here suggests that the inherent variability of CpG methylation, and possibly other molecular phenotypes, is likely to play a broad role in human biology and disease.

Chapter 4, in full, is a reprint of the material as it appears in PLoS One 2014. Vol9. Nongluk Plongthongkum, Kristel R. van Eijk, Simone de Jong, Tina Wang, Jae Hoon Sul, Marco P.M. Boks, Rene S. Kahn, Ho-Lim Fung, Roel A. Ophoff, and Kun Zhang. Characterization of Genome-Methylome Interactions in 22 Nuclear Pedigrees. PLoS One 9(7), (2014):e99313. The dissertation author was the primary investigator and author of this paper.

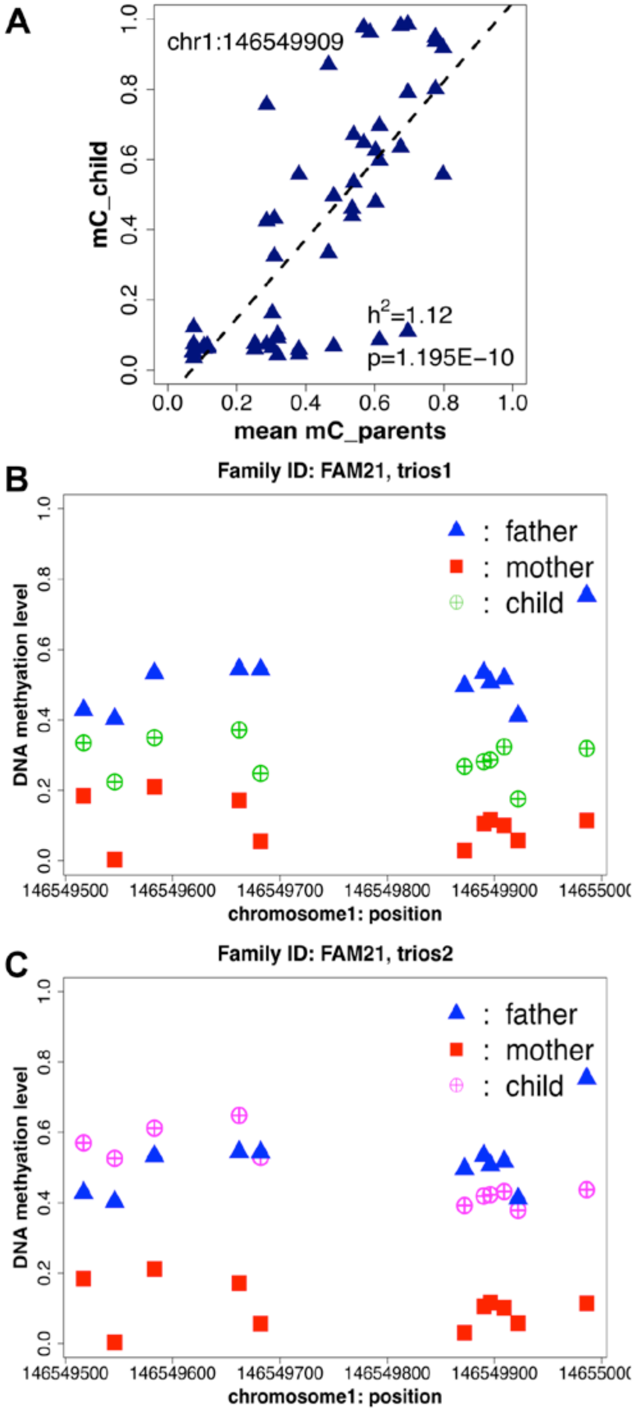


Figure 4.1 Identification of heritable CpG methylation by mid-parent offspring (MPO) analysis. (a) An example of mid-parent offspring regression of DNA methylation at the CpG site chr1:146549909. (b,c) DNA methylation level of heritable CpG at chr1:146549909 and the adjacent heritable CpGs on the same cluster exhibiting consistent pattern of DNA methylation between parents and their offspring on the two trios from the same family

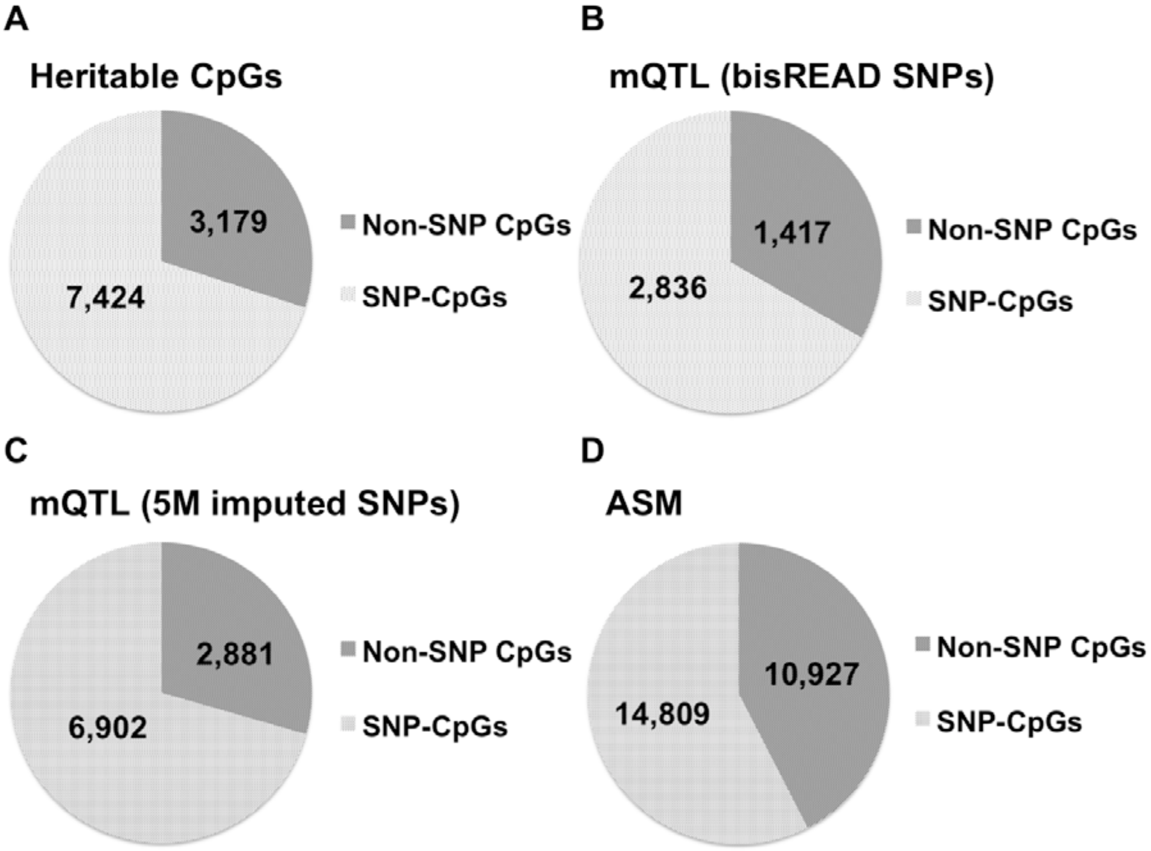


Figure 4.2 Fraction of non-SNP CpGs and SNP-CpG identified in MPO, mQTL, and ASM analysis. (a) Pie chart showing the number of heritable non-SNP CpGs and heritable SNP-CpGs. (b, c) Pie charts showing the fraction of mQTL associated non-SNP CpG and SNP-CpGs from mQTL analysis using bisREAD SNP data and 5M imputed SNP array data, respectively. (d) Pie chart showing the fraction of non-SNP CpG ASM and SNP-CpG ASM exist in at least one subject.

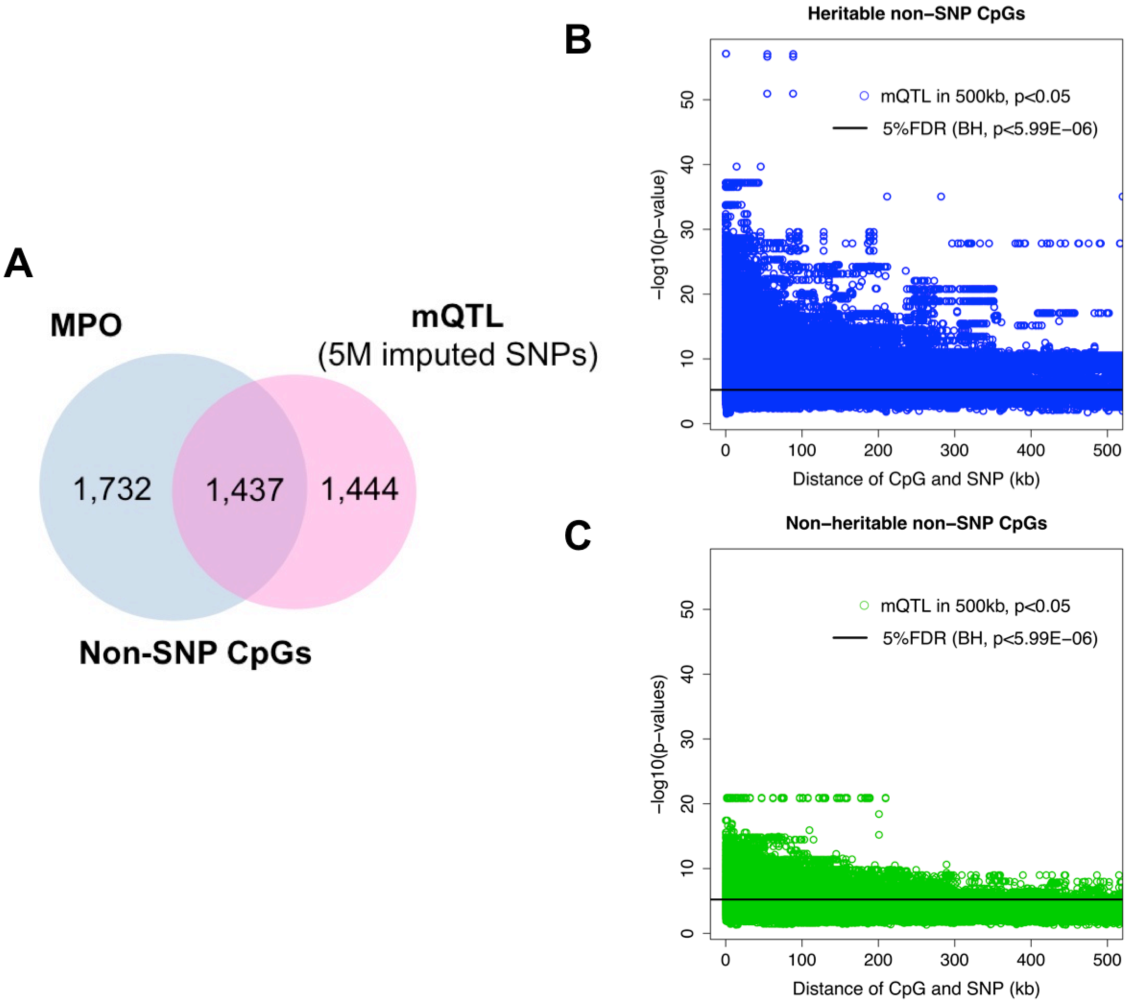


Figure 4.3 Mapping of CpG sites identified in MPO and mQTL analyses. (a) Venn diagrams showing overlap between non-SNP CpG sites significant in mQTL on 5,257,772 imputed SNPs and heritable CpGs. (b, c) Distribution of heritable CpGs and non-heritable CpGs and associated SNP pair distance within 500kb and their corresponding p-values from mQTL analysis on imputed SNPs.

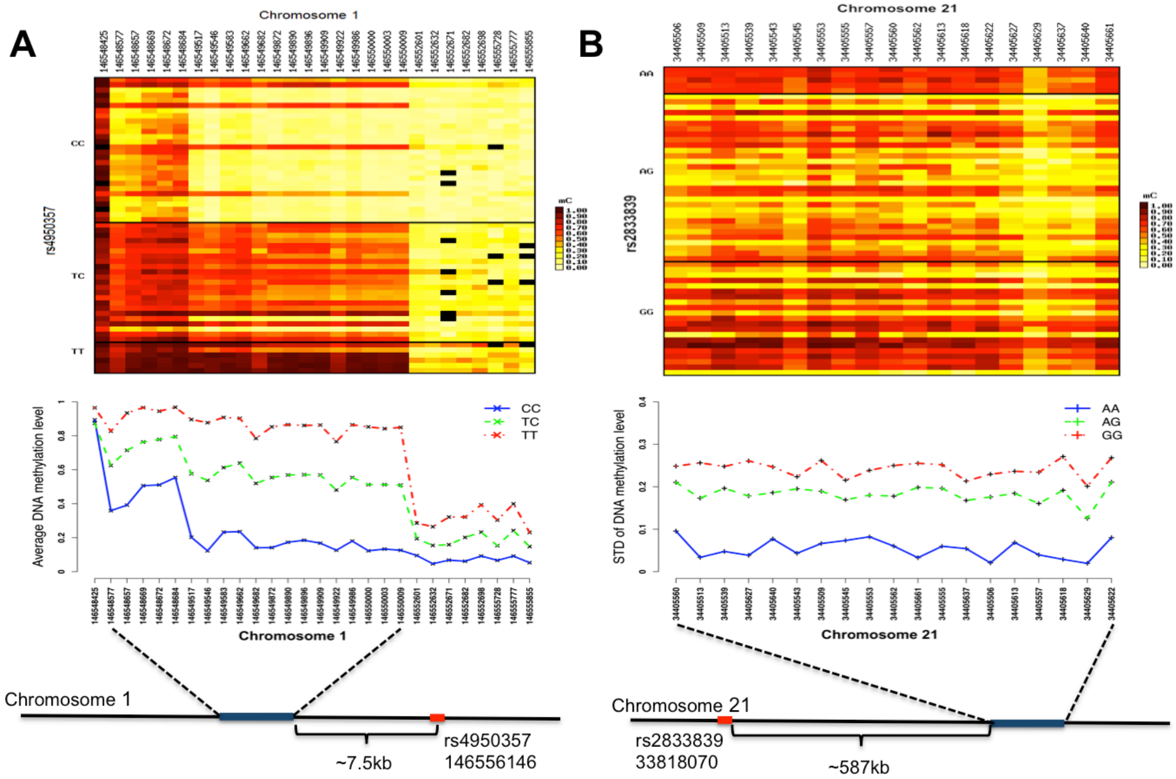


Figure 4.4 Genotype effects on the mean and variance of DNA methylation (a) Heatmap and line plot showing the association between rs4950357 SNP and the mean methylation of heritable CpGs cluster on chromosome 1 (chr1: 146548425-146555855). (b) The association of rs2833839 vSNP and the variance of methylation on VMR (chr21:34405506-34405661).

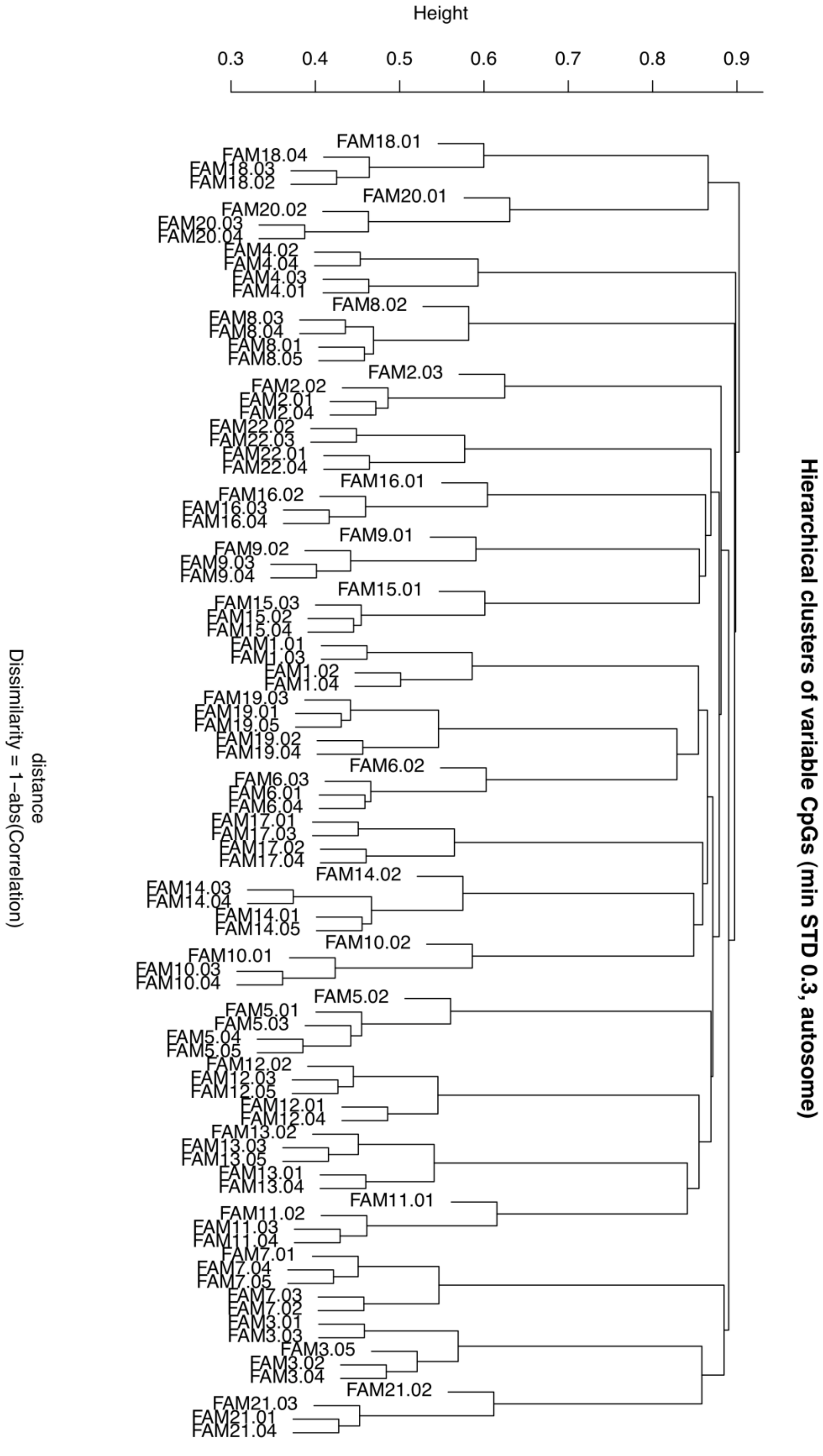


Figure 4.S1 Hierarchical clustering of high variable CpGs.

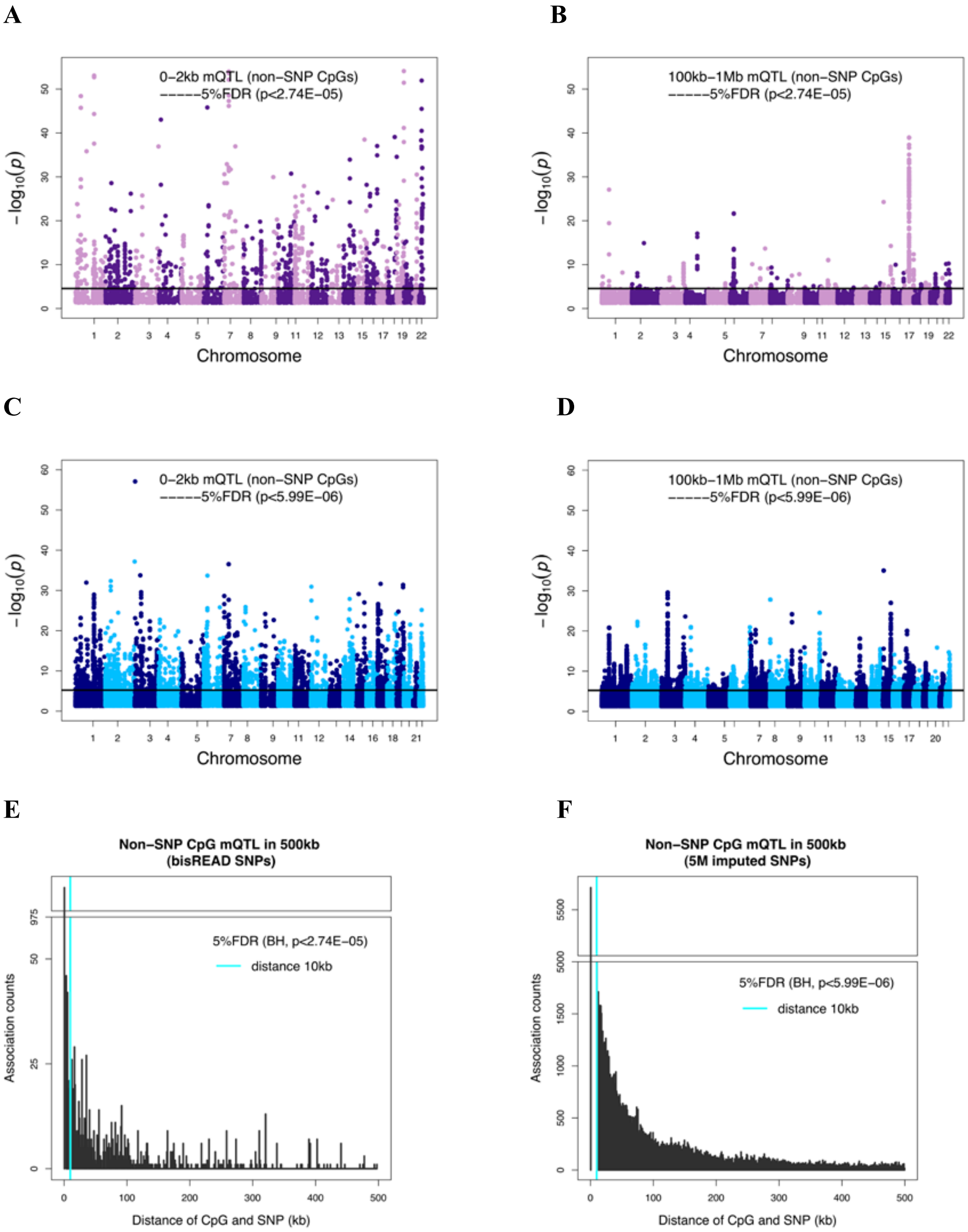


Figure 4.S2 Manhattan and density plots showing the distribution of associated CpG and SNP pairs across all chromosomes between CpG and SNP pair of 0-2kb (left) and 100kb-1Mb (right) of mQTL analysis using bisREAD SNP data (a, b) and 5M imputed SNP data (c, d), respectively. Distribution of CpG and SNP associations and their corresponding absolute distances of mQTL analysis using bisREAD SNP data (e) and 5M imputed SNP data (f), respectively.

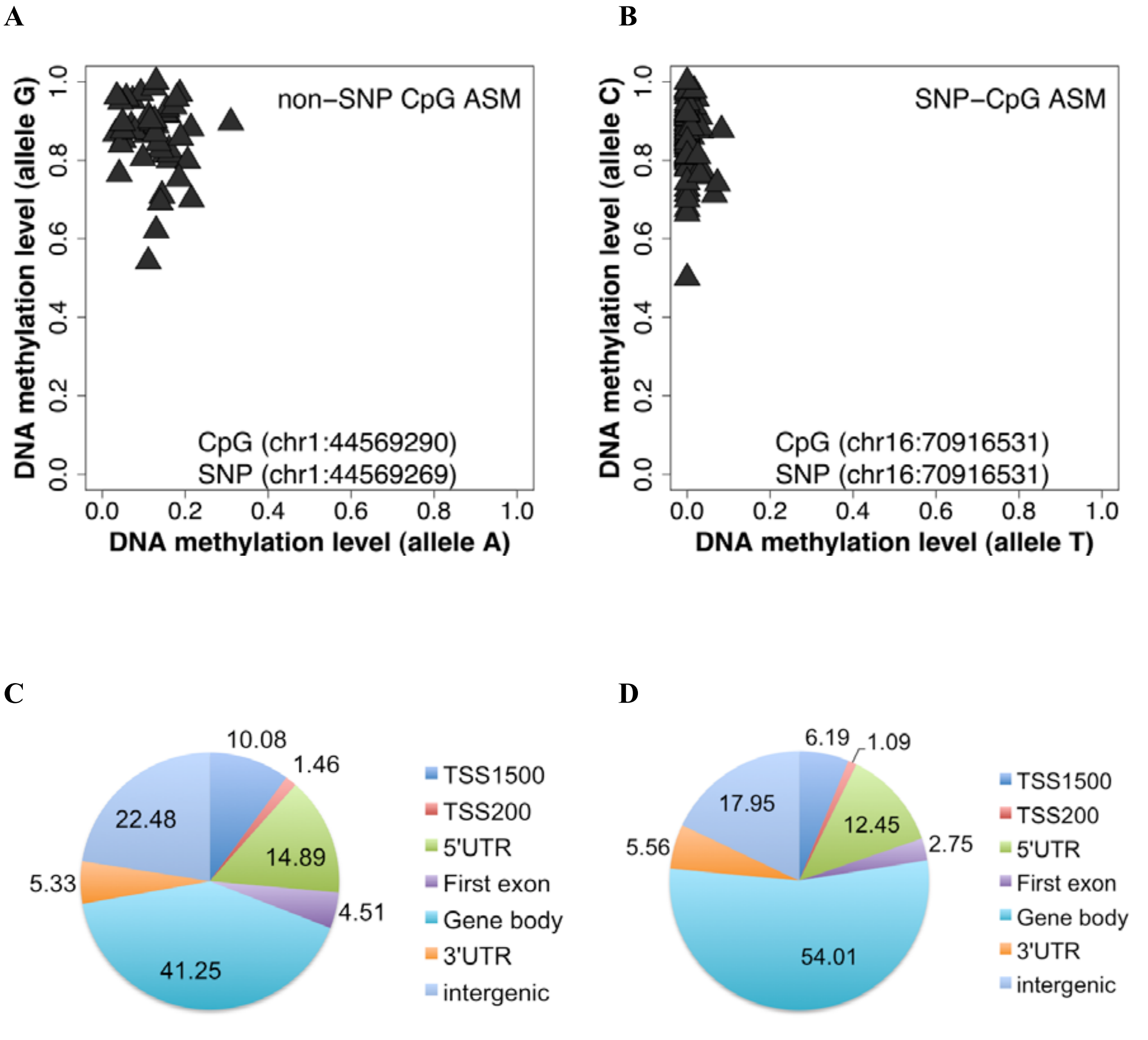


Figure 4.S3 Examples of ASM events and regional annotation of CpG associated with ASM. (a, b) Example of allele specific DNA methylation of non-SNP CpG and SNP-CpG, respectively. (b) The presence of T SNP on CpG sites disrupted DNA methylation of that allele. (c, d) Pie charts showing the distribution of non-SNP CpG ASM and SNP-CpG ASM, respectively, in different regions.

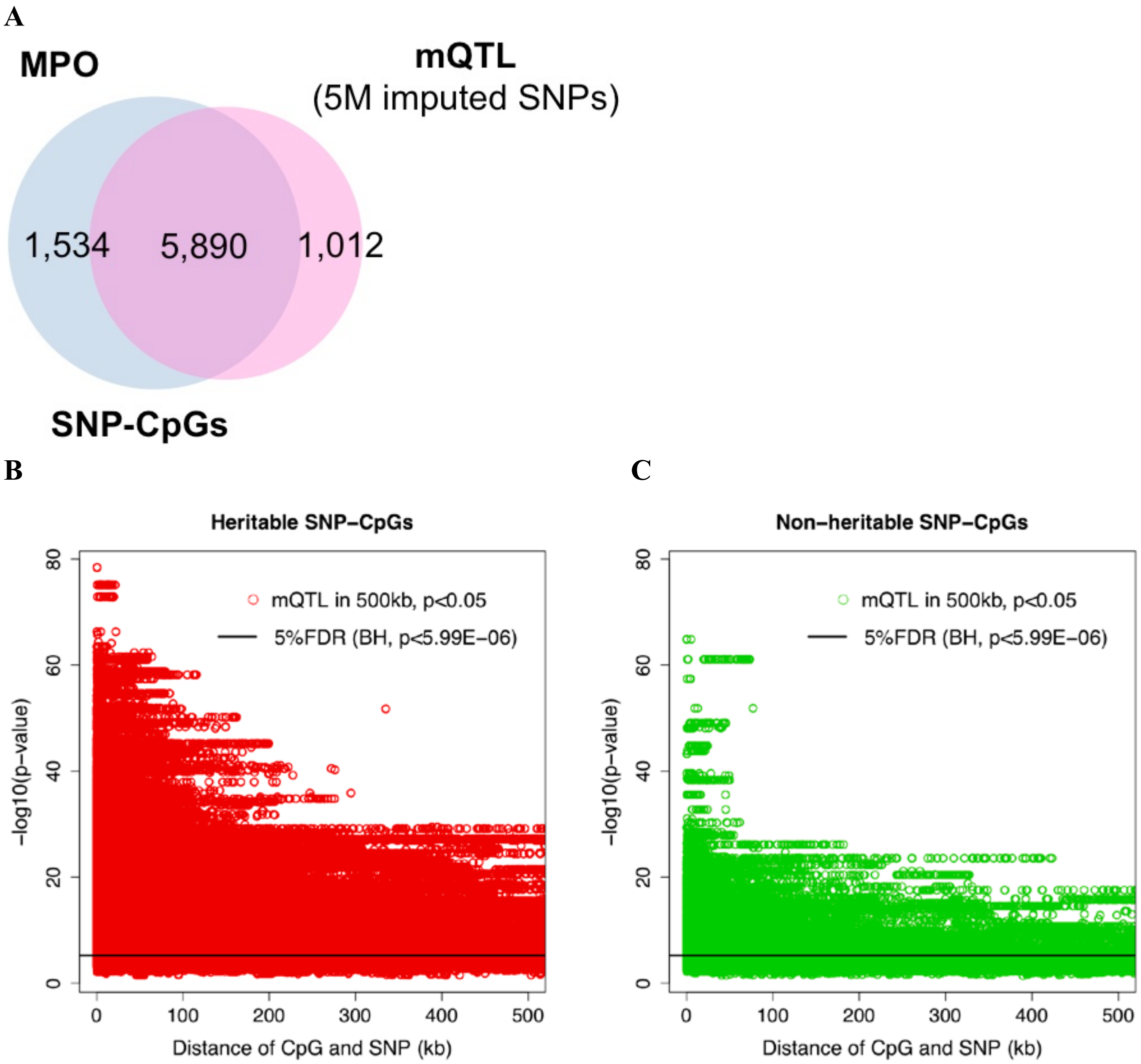


Figure 4.S4 (a) Venn diagrams showing overlap between SNP-CpG significant in mQTL and MPO analyses (based on the 5M imputed SNPs). (b, c) Distribution of heritable CpG and non- heritable CpGs, respectively, and SNP pair in mQTL analysis within 500kb and their corresponding p-values.

Table 4.1 The top 10 VMR clusters and their associated genes. The genes in bold text expressed at detectible level in whole blood and were selected for association testing.

Number of variable CpGs in VMR clusters	VMR cluster coordinates	Associated genes
53	chr20:57426730-57464571	**GNAS**, GNAS-AS1
49	chr8:144358566-144371985	GLI4, **ZNF696**
47	chr7:27143370-27184750	HOXA2, HOXA3, **HOXA5**, HOXA6, HOXA-AS3
44	chr5:140718989-140863492	PCDHGA1,PCDHGA2,PCDHGA3, PCDHGA4, **PCDHGA5**,PCDHGA6,PCDHGA7, PCDHGA8, PCDHGA11,PCDHGB1,PCDHGB2, PCDHGB3, **PCDHGB4**,PCDHGB7,PCDHGB8P, PCDHGC3, PCDHGC4
41	chr20:32255315-32255936	ACTL10,**NECAB3**
35	chr5:135415001-135416725	VTRNA2-1
28	chr19:57349099-57352134	MIMT1, **PEG3**, ZIM2
26	chr8:145162974-145164623	**KIAA1875**, **MAF1**
26	chr11:7110142-7110456	RBMXL2
24	chr1:205818899-205819600	PM20D1

Table 4.S1 Distribution of of heritable CpG sites based on genomic regions (percentage)

Methylation data	TSS1500	TSS200	5' UTR	First exon	Gene body	3' UTR	Intergenic
heritable non-SNP CpGs	9.39	2.59	15.71	5.48	42.85	5.60	18.37
heritable SNP-CpGs	5.91	0.92	12.64	3.47	53.77	5.72	17.58
variable CpGs (min STD 0.1)	10.98	1.74	13.85	5.77	44.72	5.59	17.34
all characterized CpGs	12.15	4.30	15.43	8.77	40.24	5.52	13.60

Table 4.S2 Distribution of CpG and SNP associations at different distance between CpG and SNP pairs (bisREAD SNPs).

Distance of CpG and SNP	Number of associations	% of total number of associations
0-2kb	1,071	47.6
0-10kb	1,331	59.2
10-20kb	142	6.3
20-30kb	102	4.5
30-40kb	84	3.7
40-50kb	51	2.3
0-100kb	1,907	84.8
0-150kb	1,986	88.3
100kb-1Mb	341	15.2
150kb-1Mb	262	11.7

Table 4.S3 Distribution of CpG and SNP associations at difference distance between CpG and SNP pairs (5M imputed SNPs)

Distance of CpG and SNP	Number of associations	% of total number of associations
0-2kb	9,325	6.4
0-10kb	29,052	20.1
10-20kb	15,648	10.8
20-30kb	11,512	8.0
30-40kb	8,669	6.0
40-50kb	7,003	4.8
0-100kb	93,960	64.9
0-150kb	105,711	73.0
100kb-1Mb	50,820	35.1
150kb-1Mb	39,069	27.0

Table 4.S4 Number of non-SNP CpG showing ASM shared by multiple individuals and the overlap with heritable CpGs

# of subjects	# of ASM CpGs	# of ASM CpGs found heritable	% of ASM CpGs found heritable	# of ASM CpGs found variable	% of ASM CpGs found heritable & variable
1	6,079	97	1.60	1745	5.56
2	2,005	65	3.24	775	8.39
3	918	43	4.68	385	11.17
4	458	41	8.95	208	19.71
5	297	29	9.76	148	19.59
6	216	23	10.65	114	20.18
7	148	26	17.57	84	30.95
8	104	23	22.12	58	39.66
9	91	24	26.37	56	42.86
10	70	22	31.43	49	44.90
11	56	19	33.93	39	48.72
12	39	11	28.21	27	40.74
13	37	10	27.03	24	41.67
14	43	17	39.53	28	60.71
15	35	11	31.43	26	42.31

Table 4.S5 Genomic region annotation of CpG ASM (percentage)

Methylation data	**TSS1500**	**TSS200**	**5' UTR**	**First exon**	**Gene body**	**3' UTR**	**Intergenic**
Non-SNP CpG ASM	10.08	1.46	14.89	4.51	41.25	5.33	22.48
SNP-CpG ASM	6.19	1.09	12.45	2.75	54.01	5.56	17.95
all captured CpGs	12.2	4.6	15.8	9.0	38.8	5.3	14.3

Table 4.S6 Mendel error rates of SNP genotypes

bisREAD SNPs Mendel error rate

Family IDs	CHLD	N	Mendel error rate
FAM1	2	890	0.058
FAM2	2	1944	0.126
FAM3	3	390	0.025
FAM4	2	616	0.040
FAM5	3	429	0.028
FAM6	2	345	0.022
FAM7	3	701	0.045
FAM8	3	848	0.055
FAM9	2	437	0.028
FAM10	2	204	0.013
FAM11	2	342	0.022
FAM12	3	396	0.026
FAM13	3	562	0.036
FAM14	3	1000	0.065
FAM15	2	431	0.028
FAM16	2	645	0.042
FAM17	2	300	0.019
FAM18	2	835	0.054
FAM19	3	466	0.030
FAM20	2	281	0.018
FAM21	2	233	0.015
FAM22	2	785	0.051

5M Imputed SNPs Mendel error rate

Family IDs	CHLD	N	Mendel error rate
FAM6	1	293	5.573E-05
FAM18	1	161	3.062E-05
FAM11	1	115	2.187E-05
FAM12	2	800	1.522E-04
FAM5	1	3013	5.731E-04
FAM1	2	7631	1.451E-03

CHLD: number of offspring in each family, N: number of Mendel error in each family

4.6 References

1. McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR and Birney E (2010) Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. Science 328:235-239.

2. Zhang YY, Rohde C, Reinhardt R, Voelcker-Rehage C and Jeltsch A (2009) Non-imprinted allele-specific DNA methylation on human autosomes. Genome Biology 10.

3. Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, Plomin R and Mill J (2010) Allelic Skewing of DNA Methylation Is Widespread across the Genome. American Journal of Human Genetics 86:196-212.

4. Rakyan VK, Down TA, Balding DJ and Beck S (2011) Epigenome-wide association studies for common human diseases. Nature reviews. Genetics 12:529-41.

5. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T and Zhang K (2013) Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular Cell 49:359-367.

6. Tycko B (2010) Allele-specific DNA methylation: beyond imprinting. Human molecular genetics 19:R210-20.

7. Feinberg AP (2007) Phenotypic plasticity and the epigenetics of human disease. Nature 447:433-440.

8. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y and Pritchard JK (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome biology 12:R10.

9. Fraser HB, Lam LL, Neumann SM and Kobor MS (2012) Population-specificity of human DNA methylation. Genome biology 13:R8.

10. Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, Kucera KS, Willard HF and Myers RM (2011) Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS genetics 7:e1002228.

11. van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, de Kovel CG, Janson E, Strengman E, Langfelder P, Kahn RS, van den Berg LH, Horvath S and Ophoff RA (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC genomics 13:636.

12. Hellman A and Chess A (2010) Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics & chromatin 3:11.

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

14. Fang F, Hodges E, Molaro A, Dean M, Hannon GJ and Smith AD (2012) Genomic landscape of human allele-specific DNA methylation. Proceedings of the National Academy of Sciences of the United States of America 109:7332-7337. doi:

15. Schilling E, El Chartouni C and Rehli M (2009) Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. Genome Research 19:2028-2035.

16. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y and Zhang K (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 27:353-60.

17. Diep D, Plongthongkum N, Gore A, Fung HL, Shoemaker R and Zhang K (2012) Library-free methylation sequencing with bisulfite padlock probes. Nature Methods 9:270-U69.

18. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P and Dermitzakis ET (2007) Population genomics of human gene expression. Nat Genet 39:1217-24.

19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81:559-75.

20. Browning BL and Browning SR (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. American Journal of Human Genetics 84:210-223.

21. Howie B, Fuchsberger C, Stephens M, Marchini J and Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44:955-9.

22. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME and McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-73.

23. Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

24. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB and Shen R (2011) High density DNA methylation array with single CpG site resolution. Genomics 98:288-295.

25. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK and Kelsey KT (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13:86.

26. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekstrom TJ and Feinberg AP (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nature Biotechnology 31:142-147.

27. Jaffe AE and Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome biology 15:R31.

28. Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL and Ren B (2012) Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. Cell 148:816-831.

29. Greaves I, Groszmann M, Dennis ES and Peacock WJ (2012) Trans-chromosomal methylation. Epigenetics : official journal of the DNA Methylation Society 7:800-5.

30. Gimelbrant A, Hutchinson JN, Thompson BR and Chess A (2007) Widespread monoallelic expression on human autosomes. Science 318:1136-1140.

31. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA and Feinberg AP (2011) Increased methylation variation in epigenetic domains across cancer types. Nature Genetics 43:768-U77.

32. Feinberg AP and Irizarry RA (2010) Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proceedings of the National Academy of Sciences of the United States of America 107 Suppl 1:1757-64.

33. Bastepe M (2007) The GNAS Locus: Quintessential Complex Gene Encoding Gsalpha, XLalphas, and other Imprinted Transcripts. Current genomics 8:398-414.

34. Plagge A and Kelsey G (2006) Imprinting the Gnas locus. Cytogenetic and genome research 113:178-87.

35. Luykx JJ, Bakker SC, Lentjes E, Neeleman M, Strengman E, Mentink L, Deyoung J, de Jong S, Sul JH, Eskin E, van Eijk K, van Setten J, Buizer-Voskamp JE, Cantor RM, Lu A, van Amerongen M, van Dongen EP, Keijzers P, Kappen T, Borgdorff P, Bruins P, Derks EM, Kahn RS and Ophoff RA (2013) Genome-wide association study of monoamine metabolite levels in human cerebrospinal fluid. Molecular psychiatry.

36. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y and Pritchard JK (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482:390-394.

37. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y and Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768-772.

38. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES and Liu C (2010) Genetic control of individual differences in gene-specific methylation in human brain. American journal of human genetics 86:411-9.

39. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE and Stefansson K (2008) Genetics of gene expression and its effect on disease. Nature 452:423-8.

40. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, Holmes L, Meldrim J, Kelley F, Gnirke A and Meissner A (2012) Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome biology 13:R92.

41. Teperino R, Lempradl A and Pospisilik JA (2013) Bridging epigenomics and complex disease: the basics. Cellular and molecular life sciences : CMLS.

42. Drong AW, Nicholson G, Hedman AK, Meduri E, Grundberg E, Small KS, Shin SY, Bell JT, Karpe F, Soranzo N, Spector TD, McCarthy MI, Deloukas P, Rantalainen M and Lindgren CM (2013) The presence of methylation quantitative trait Loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One 8:e55923.

43. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, Belvisi MG, Brown R, Vineis P and Flanagan JM (2013) Epigenome-wide

association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Human Molecular Genetics 22:843-851.

44. Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, Huang Z, Hoyo C, Midttun O, Cupul-Uicab LA, Ueland PM, Wu MC, Nystad W, Bell DA, Peddada SD and London SJ (2012) 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environmental health perspectives 120:1425-31.

45. Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ and Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nature biotechnology 27:361-8.

46. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews. Genetics 13:484-92.

47. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B and Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315-322.

48. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK and Schubeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480:490-5.

49. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kelllis M, Kheradpour P, Lassman T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SC, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LA, Adams LB, Kelly CJ, Zhang J, Wexler JR, Good PJ, Feingold EA, Crawford GE, Dekker J, Elinitski L, Farnham PJ, Giddings MC, Gingeras TR, Guigo R, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Starnatoyannopoulos JA, Tennebaum SA, Weng Z, White KP, Wold B, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Eaton ML, Dobin A, Lassmann T, Tanzer A, Lagarde J, Lin W, Xue C, Williams BA, Zaleski C, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac

S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Robyr D, Ruan X, Sammeth M, Sandu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Hayashizaki Y, Reymond A, Antonarakis SE, Hannon GJ, Ruan Y, Carninci P, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Grasfeder LL, Giresi PG, Battenhouse A, Sheffield NC, Showers KA, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Iyer VR, Sandhu KS, Zheng M, Wang P, Gertz J, Vielmetter J, Partridge EC, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, Muratet MA, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Newberry JS, Levy SE, Absher DM, Wong WH, Blow MJ, Visel A, Pennachio LA, Elnitski L, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Davidson C, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Hunt T, Jungreis I, Kay M, Khurana E, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tapanari E, Tress ML, van Baren MJ, Washieti S, Wilming L, Zadissa A, Zhengdong Z, Brent M, Haussler D, Valencia A, Raymond A, Addleman N, Alexander RP, Auerbach RK, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyenger S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Larnarre-Vincent N, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Raha D, Ramirez L, Reed B, Shi M, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Struhl K, Weissman SM, Tenebaum SA, Penalva LO, Karmakar S, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Victorsen A, Auer T, Centarin L, Eichenlaub M, Gruhl F, Heerman S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Jain G, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Johnson AK, Johnson EM, Kutyavin TM, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sanchez ME, Sandstrom RS, Shafer AO, Stergachis AB, Thomas S, Vernot B, Vierstra J, Vong S, Weaver MA, Yan Y, Zhang M, Akey JA, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Stamatoyannopoulos JA, Beal K, Brazma A, Flicek P, Johnson N, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Miller W, Bickel PJ, Banfai B, Boley NP, Huang H, Li JJ, Noble WS, Bilmes JA, Buske OJ, Sahu AO, Kharchenko PV, Park PJ, Baker D, Taylor J and Lochovsky L (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

50. Liu Y, Siegmund KD, Laird PW and Berman BP (2012) Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome biology 13:R61.

51. Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. Nature 402:C47-52.

52. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U (2002) Network motifs: Simple building blocks of complex networks. Science 298:824-827.

53. Becskei A and Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405:590-593.

54. Raj A, Rifkin SA, Andersen E and van Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. Nature 463:913-U84.

CHAPTER 5 CONCLUSIONS

The goals of this dissertation are (1) to develop a high accuracy and throughput method for targeted quantification of DNA methylation, (2) to validate the performance of the improved bisulfite padlock probe (BSPP) method to be implement in clinical diagnostics for quantification of locus-specific DNA methylation biomarker, and (3) to investigate the influence of genetic polymorphisms on variability of DNA methylation. In chapter 2 of this dissertation, I described the improvement of BSPP method. Zhang's lab firstly have developed a program called ppDesigner to design a high efficient bisulfite padlock probes. To validate the assay, we used ppDesigner to design a genome-scale probe set containing ~330,000 probes to capture selective CpG sites on human genome. I used this probe set that we have normalized the probe efficiency by subsetting and using suppressor oligonucleotides to analyze H1 embryonic stem cells (H1 ESCs) and PGP1 fibroblast. With this probe set, I was able to capture ~480,904 CpG sites on average. I have demonstrated that the data generated by BSPPs accurately represented the methylation status of the selective targets. There were the consistencies of methylation data within the same batch and between different batches (Pearson's correlation coefficient R = 0.97-0.98) and also between the technical replicates. I also showed that H1 ESC methylation data generated by BSPPs were consistent with the published whole-genome bisulfite sequencing (WGBS) data (Pearson's correlation coefficient R = 0.95). Another improvement of BSPP is the implement of library-free approach by skipping the regular steps of shotgun library preparation. I used multiplexed primers with 6-base pair (bp) barcodes to directly amplify the captured DNA. This feature allowed me to routinely generate 96 individual libraries and sequenced all in the same sequencing run. Zhang's lab has also built the bioinformatics pipeline for read mapping and methylation quantification, called bisReadMapper. The pipeline is compatible to data generated by targeted and whole-genome bisulfite sequencing. I was also able to call SNPs simultaneously with methylation mapping, which allowed me to be able to track the samples, which is useful for projects handling large sample sizes.

In chapter 3, I extensively validated the performance of our BSPP developed in chapter 2 for implementing in clinical diagnostic for routinely methylation biomarker analysis. I have performed BSPP capture in parallel with other research groups that performed their assays for comparison on the same sample set. The study was designed to asses the values of the assays, including accuracy, sensitivity, specificity, throughput, easy workflow, and cost. From the 1st report of the technology comparison, I have shown that BSPP is among the assays that had a good performance in average based on accuracy and consistency to other assays. BSPP has also showed a high throughput, which is a strength feature of this assay.

In chapter 4, my aim is to investigate the regulation of DNA methylation level by genetic variances. I applied the DMR330k probe set as described in chapter 2 to characterize DNA methylation status of 96 samples from 22 nuclear pedigrees consisting of 52 trios. In this study, I took the advantage of the samples with family structure to assess the full extend of heritable CpG sites by mid-parent offspring (MPO) analysis. We have identified 10,593 heritable CpG sites, and we found that 70% of the heritable CpGs were the SNPs that present on CpG sites. I have used the two independent approaches including methylation quantitative trait loci (mQTL) and allele-specific DNA methylation (ASM) analysis to identify the *cis*-regulatory SNPs associated with heritable

CpGs. I have demonstrated that *cis*-regulatory SNPs identified by mQTL analysis accounted for only roughly half of the heritable CpG methylation, whereas ASM analysis was only able to identify 5% of *cis*-regulatory SNPs. These results showed that the full extend of *cis*-regulatory SNPs associated with heritable CpGs was not able to identified by mQTL analysis, and ASM analysis is far less powerful than mQTL analysis. Finally, I have identified SNPs associated with the variance of multiple nearby CpG sites. This finding supported the recently purposed hypothesis by Feinberg's group that genetic variants are not only associated with the mean but also the variance of molecular phenotypes such as DNA methylation or gene expression. Overall, in this chapter, I have shown that the extent of genome-methylome interactions is well beyond what is detectible with the commonly used mQTL and ASM analysis

In summary, I have developed targeted bisulfite sequencing technique or BSPP that has a high accuracy and throughput and is scalable to be applied in a wide-range of applications. Using BSPPs I can characterize DNA methylation status in genome-wide scale or in small target sizes for a board-range of applications such as methylation biomarker detection or detection of DNA methylation aberration at selective regions. I also used BSPPs to characterize the effects of genetic polymorphisms on the mean and variability of DNA methylation.