UNIVERSITY OF CALIFORNIA, SAN DIEGO

Inter- and Intracellular Dynamics of DNA Methylation

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

in

Chemistry

by

Robert Field Shoemaker

Committee in charge:

Professor Wei Wang, Chair

Professor Alexander Hoffman

Professor Gourisankar Ghosh

Professor Bing Ren

Professor Peter Wolynes

2010

The dissertation of Robert Field Shoemaker is approved, and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

ACKNOWLEDGEMENTS

I would like to thank Professor Wei Wang for his patience and guidance as we endeavored to learn something new on the research frontier.

I would also like to thank Professor Kun Zhang for his generous collaboration. The contributions from his lab made this dissertation possible.

I would finally like to thank my parents for their unwavering support.

Chapter 1, in full, is a reprint of the material as it appears in R. Shoemaker, W. Wang, K. Zhang. DNA methylation and its mediators. *Wiley Interdiscip Rev Syst Biol Med*, Epub (Sep, 2010). The dissertation author was the primary author of this paper.

Chapter 3 is a rewrite of material as it appears in J. Deng, R. Shoemaker, B. Xie, A. Gore, E. M. LeProust, J. Antosiewicz-Bourget, D. Egli, N. Maherali, I. H. Park, J. Yu, G. Q. Daley, K. Eggan, K. Hochedlinger, J. Thomson, W. Wang, Y. Gao, K. Zhang, Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353 (Apr, 2009). The dissertation author was the primary data analyst of this paper.

Chapter 4, in part, is a reprint of the material as it appears in R. Shoemaker, J. Deng, W. Wang, K. Zhang, Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome*

Res **20**, 883 (Jul, 2010). The dissertation author was the data analyst and one of the principal authors of this paper.

VITA

2004 Bachelor of Arts, University of California, San Diego

2004 Bachelor of Science, University of California, San Diego

2006 Master of Science, University of California, San Diego

2010 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E. M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I. H., Yu, J., Daley, G. Q., Eggan, K., Hochedlinger, K., Thomson, J., Wang, W., Gao, Y. & Zhang, K. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353-360, doi:nbt.1530 10.1038/nbt.1530 (2009).

Won KJ, Agarwal S., Shen L., Shoemaker R. Ren B., Wang W. An integrated approach to identifying cis-regulatory modules in the human genome*. PLoS One* **4**, e5501, doi:10.1371/journal.pone.0005501 (2009).

Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* **20**, 883-889, doi:gr.104695.10910.1101/gr.104695.109 (2010).

Shoemaker, R, Wang W., Zhang, K. Mediators and Dynamics of DNA Methylation. *Wiley Interdiscip Rev Syst Biol Med*, Epub (2010).

ABSTRACT OF THE DISSERTATION

Inter- and Intracellular Dynamics of DNA Methylation

by

Robert Field Shoemaker

Doctor of Philosophy in Chemistry

University of California, San Diego, 2010

Professor Wei Wang, Chair

CpG methylation in the human genome plays an important role in the transcriptional regulatory network. I have explored the biology of this epigenetic mark and I have developed tools that aid in methylation-related analyses. Although most of the human genome is methylated, regions involved in regulatory activity tend to be unmethylated. The methylation states

of CpGs influence DNA-protein interactions. Since DNA methylation influences transcriptional regulation, DNA methylation signatures discriminate among phenotypes, such as cancerous cells versus healthy cells or differentiated cells versus pluripotent cells.

To better study DNA methylation, I produced a new probe set, Cpg220k, used in targeted bisulfite sequencing by Kun Zhang's lab. This involved surveying the literature for experiments that found regions of variable DNA methylation and using other experimental data, such as DNaseI hypersensitivity data, that found candidates for likely differential methylation. I produced padlock probes based on maximizing the bp coverage of these regions while minimizing the cost of the experiment.

Using targeted bisulfite sequencing, I developed analytical tools that allowed for inter- and intracellular analysis of CpG methylation data. I found that methylation signatures accurately classify ES, fibroblast, and iPS cell lines. Gene expression and methylation are negatively correlated at the TSS but they are weakly correlated further down- and upstream from the TSS. Further exploring fuzzily methylated CpGs (methylation frequencies between .25 and .75), I found regions that exhibited allele-specific methylation. Many of the ASM regions involved a SNP overlapping with a CpG site, thereby creating sequence dependent ASM. Other ASM regions were likely products of biological regulatory mechanisms.

I then created a pipeline that utilized the tools I developed for my methylation analyses. I integrated the pipeline into a website (http://wanglab.ucsd.edu/star/) so that other scientists can conveniently run these analyses on their methylation data. Due to the increasing scale of methylation experiments, I created the pipeline with the capability of handling massive data sets that can exceed several billion reads. I made the tools that were critical to my research publicly available in hopes to furthering the scientific community's understanding of DNA methylation.

Chapter 1 Introduction to DNA Methylation

1.1 Introduction Overview

DNA methylation is an inherited epigenetic chemical modification that occurs primarily on the cytosines in CpG dinucleotides. However, examples of non-CpG methylation are found in plants and human embryonic stem cells(*1-3*). The proteins DNMT1 and DNMT3a and DNMT3B are known to methylate cytosines while protein families including MBD, SRA, and other zinc finger proteins recognize the presence of methylated cytosines(*4-13*). DNMT1 maintains methylation patterns and this maintenance is required for normal cell division(*14-16*). Originally thought to bind DNA due to its similar structure and sequence to other members within the DNMT family, DNMT2 was later found to specifically target RNA(*14, 17, 18*). DNMT3A and DNMT3B have de novo methylation activity that is required for embryonic development(*19*). Another protein, DNMT3L, co-localizes with DNMT3A and DNMT3B and it enhances *de novo* methylation *in vitro* and *in vivo*(*20-26*). DNMT3L itself lacks methyltransferase activity and thusly enhances methylation via its interaction with DNMT3A. DNA methylation is widespread across many organisms, including bacteria, insects, and mammals, and it is most commonly associated with transcriptional silencing(*27-30*). As many repetitive regions and transposons are methylated in various genomes, DNA methylation is thought to act as a genomic defense mechanism that prevents the activation of these

sequences(*31-34*). DNA methylation is also present in genes and regulatory regions, where it is correlated with transcriptional repression or activation depending on its context. Given the regulatory effects associated with DNA methylation, characterizing the methylation states of cytosines genome-wide reveals the biological state of the host cell on a global scale(*1-3, 35-37*). This introduction will discuss: (1) the biological meaning of DNA methylation patterns on intercellular and intracellular levels, (2) explore the importance of the mechanisms that erase these patterns, (3) discuss the role of transcription factors in DNA methylation, and (4) review the presence of a new cytosine modification, hydroxymethylcytosine, which is a product of DNA methylation.

1.2 DNA Methylation In Higher Eukaryotes

A comparison of DNA methylation patterns across 8 species revealed many common features but also striking differences (Arabidopsis, green algae, rice, poplar, honey bee, mouse, sea squirt, and zebrafish)(*27*). Repetitive regions and transposons were enriched for CpG methylation relative to nearby regions for all 8 organisms but these methylation patterns in sea squirt were much weaker. Regions of high CpG density, which are known as CpG islands (CGIs), are mostly unmethylated in vertebrates (i.e. zebrafish and mouse) but overall vertebrate methylation levels are high (70-80% global CpG methylation). Non-CpG (CHG and CHH) methylation was found in all three flowering plants; CHG and CHH methylation was enriched at transposons and repetitive regions. CHH methylation describes a methylated cytosine followed

by two nucleotides that may not be guanine. CHG methylation entails a methylated cytosine that precedes an adenine, thymine, or cytosine, followed by guanine. Unlike CG or CHG methylation, CHH sites are strand specific, which means, for example, a CHH site on the Watson strand is not found on the reverse complementary Crick strand. Though exhibiting non-CpG methylation, green algae's CHH and CHG patterns showed no enrichment in any particular genomic region. The remaining organisms displayed a much lower percentage of non-CpG methylation. Vertebrates did not show a significant enrichment of CpG methylation in exons relative to introns as was seen in all three flowering plants. The honeybee, though displaying low levels of global CpG methylation, displayed significant CpG methylation in exon regions. Green algae showed very low global CpG methylation but exons were more methylated. Sea squirt also showed preferential methylation of exon regions. This conservation study shows that DNA methylation is present in higher eukaryotes and that DNA methylation often targets certain genomic regions. However, the targeted regions are not entirely consistent across all of the 8 studied organisms.

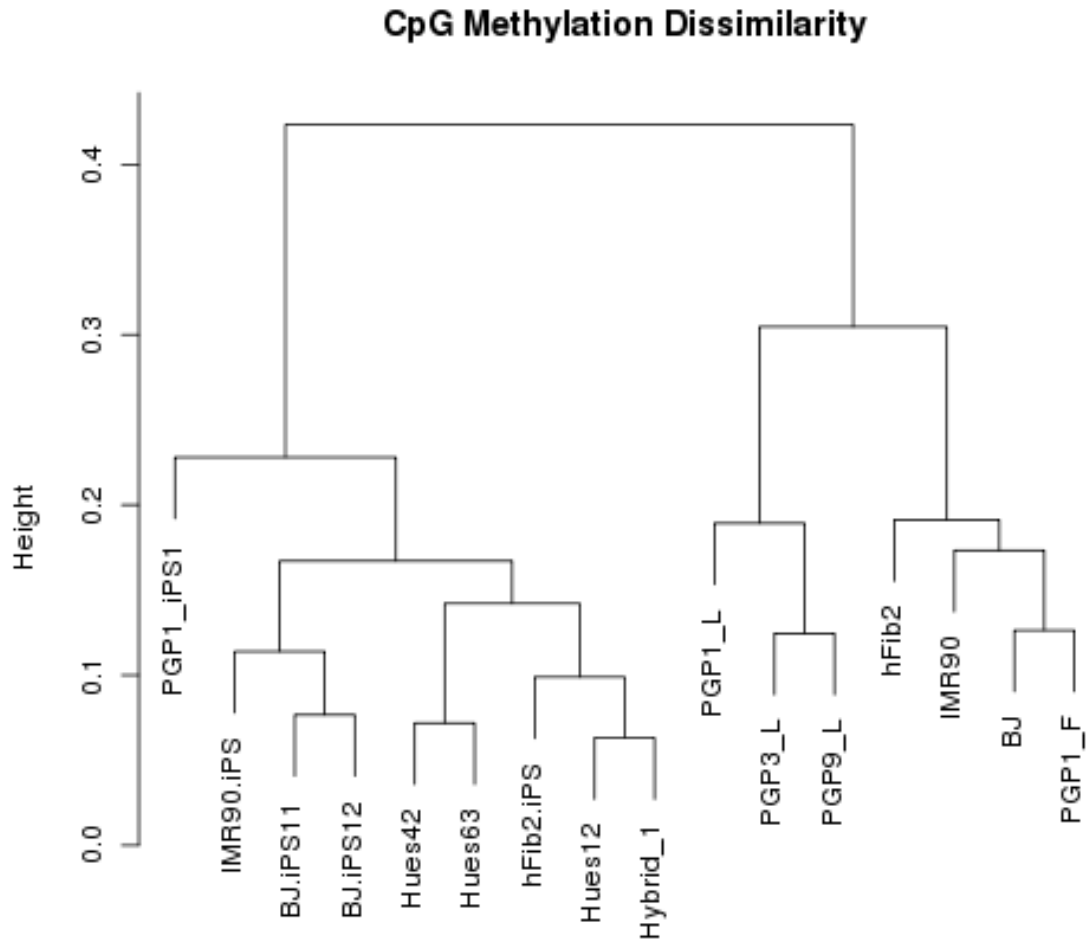


Figure 1.2.1 This figure shows the dissimilarity of cell types based on their methylation frequencies from targeted bisulfite data that covered CpG islands on chr12 and chr20. Cell types that are closer together share a more similar methylation signature. The fibroblasts, PGP1F (Personal Genome Foundation 1 Fibroblast), BJ, IMR90, and hFib2 (human Fibroblast), cluster closely together while the lymphoblasts PGP9L (Personal Genome Project 9 Lymphoblast), PGP3L, and PGP1L also cluster together closely. Regarding the pluripotent cell lines, the ES cells (HUES12, HUES42, and HUES63) cluster very closely together. The Hybrid1 cell line, which consists of fused nuclei from HUES6 and BJ, also clusters closely with the ES cell group. The iPS lines appear to be much more similar to the ES cell group than the differentiated fibroblasts and lymphoblasts but the iPS group exhibits a wider spectrum of methylation signatures than the ES cell group or the two differentiate cell groups.

Since CpG methylation sites are known in a sequenced genome, one can easily compare methylation signatures across intraspecies cell lines. Genome-wide DNA methylation patterns are cell-type specific. Our group performed targeted bisulfite sequencing of CpG islands across chr12 and chr20 and found that the methylation frequency data clearly distinguished between human embryonic stem cell lines, fibroblasts, and lymphoblast cell lines (Figure 1)(*38, 39*). An additional study reported a conserved set of regions that are uniquely methylated between human liver, spleen and brain tissues(*40*). These differentially methylated regions tended to be located just outside of CpG islands and were thusly called CpG island shores. Methylation of these CpG island shores had a strong inverse relationship with the expression of associated genes. The authors labeled 16,379 regions as differentially methylated across examined tissues (T-DMRs) and the median length of these regions was 255 bp. Extending this analysis to 13 colorectal cancer cell samples with matched normal mucosal samples, the authors found a separate set of differentially methylated regions (C-DMRs) that showed significant differences in methylation between the normal mucosal and the matched cancer samples. 45% of the 2,707 identified C-DMRs overlapped with T-DMRs (p-value < 10^{-14}), which showed that these many of these DMRs could discriminate between both tissues and colorectal cancer. A continuation of this study found 4,401 regions were differentially methylated in iPS cells relative to the untransformed fibroblast cells (R-DMRs)(*41*). These R-DMRs, similar to the C-DMRs and T-DMRs, overlapped significantly with CpG island

shore regions (over 70%). Many R-DMRs also overlapped with T-DMRs (56%). These studies showed that DNA methylation signatures are unique across many different types of cell lines.

1.3 DNA Methylation as a Clinical Biomarker

DNA methylation can classify cell types into subpopulations, which can exhibit unique phenotypes. One cancer study examined the methylation profiles of blast cells taken from 344 diagnosed acute myeloid patients(*42*). Clustering these methylation profiles created 16 unique AML sub-type clusters. Three of those patient clusters were defined by the WHO classification(*43*), eight were enriched for specific genetic or epigenetic lesions, and the remaining 5 could not be explained by current knowledge; all of these subtypes were distinct when compared to normal bone marrow cells. The authors used the methylation signatures of 18 methylation probe sets that covered 15 genes and developed a classifier that predicted the overall survival and event-free survival of an AML patient (p-value < .001, multivariate Cox proportional hazards model). These authors showed that DNA methylation signatures can act as biomarkers that foretell patients' clinical outcomes. Another study focused on breast cancer and used methylation signatures to distinguish between the different breast cancer mutation types (BRCA1, BRCA2, BRCAx)(*44*). The authors report that DNA methylation profiling predicted BRCA1, BRCA2, and BRCAx tumors with error rates of 11%, 31%, and 36%, respectively. Classification based on DNA methylation signatures

was significantly more accurate than using gene expression data, which resulted in error rates of 11%, 44%, and 71%, respectively. However, the gene expression data was able to cluster the breast cancer samples into intrinsic subtypes (i.e. basal, luminal A, luminal B, HER2-amplified, and normal-like) while the DNA methylation data could not(*45*). Another group used the DNA methylation signatures from over 300 peripheral blood samples as an indicator for ovarian cancer(*46*). Methylation studies involving colorectal cancer and breast cancer have shown that peripheral blood samples can indicate the presence of cancer when compared to healthy controls(*47-51*). Starting with 25,642 CpG sites, the authors found that a 100 CpG sites in peripheral blood cells that accurately discriminated between healthy and pre-treatment ovarian cancer patients. Using these selected 100 CpG sites, the authors were able to correctly identify ovarian cancer samples (58 healthy samples and 43 pre-treatment cases) with an AUC of .82. AUC, which stands for Area Under the (ROC) curve, describes the ability of a test to detect true positives versus false positives. Tests with AUC values close to 1 represent high true positive and low false positive detection rates and AUC values close to .5 represent discriminatory power that is similar to random selection. Tests with AUC less than .5 perform worse than random selection. Comparing a sample post-treatment population that showed signs of active disease with one that served as a healthy control, the classifier correctly identified those active disease patients with an AUC of .76. Finally the classifier was able to discriminate between post-treatment samples showing active disease and post-treatment

samples without active disease (.74 AUC). Additional tests showed the CpGs used in the classifier were not dependent on age. DNA methylation can be used to predict cancer-related phenotypes as seen in these three studies.

1.4 Allele Specific Methylation

A mammalian cell contains a paternal and a maternal chromosome set. Most biological activity is thought to be symmetrical across the paternal and maternal chromosomes but there are regions where each parental allele is regulated uniquely. Allele-specific methylation, ASM, (Figure 2a) is a common feature of these regions and ASM can be coupled with allele-specific expression, ASE (Figure 1.2.1). Imprinted genes consist of genes exhibiting ASM and ASE in a parent-of-original specific manner. ASM near the loci of these ASE genes reveals significant methylation differences where one allele is significantly more methylated than the other. There are between 50 and 100 known imprinted genes in mouse and human(*52-54*). However, over 1,000 genes have been found to exhibit parent-of-origin allelic gene expression effects in the mouse brain(*55*). Besides imprinting, ASM can also occur in a manner that is independent of parental origin. This occurs in chromosome X inactivation randomly but it has also been unexpectedly observed in numerous autosomal regions(*39, 56*). This section explores ASM and places it in the context of other regulatory mechanisms that co-occur at these sites.

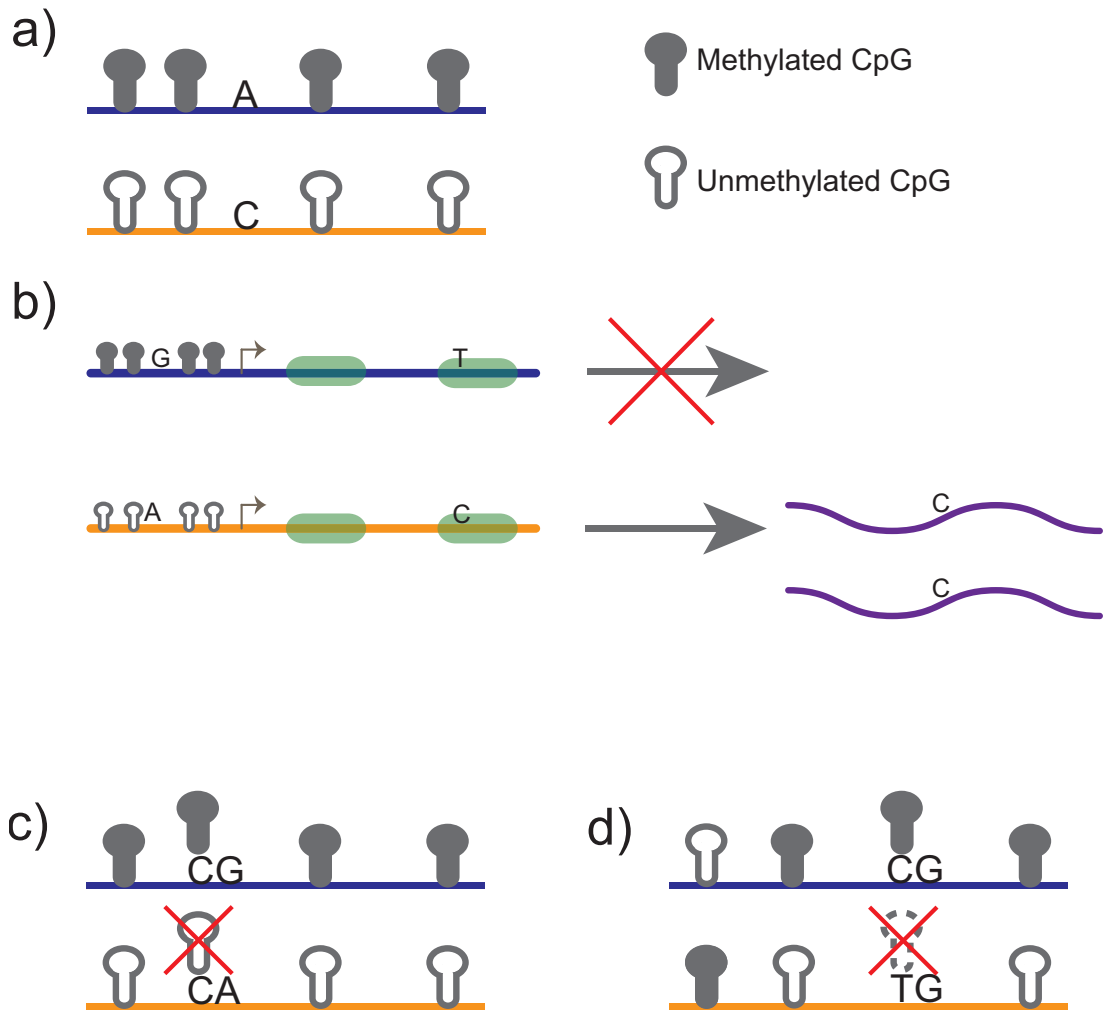


Figure 1.4.1 a) A schematic of typical ASM. The two alleles are represented separately as blue and orange lines. A SNP distinguishes the alleles and the blue allele is methylated while the orange allele is not. This is considered an ASM region. b) This figure shows a gene whose promoter region is methylated in the blue allele and unmethylated in the orange allele. A SNP site at the promoter distinguishes the alleles. Additionally, another SNP site in an exon differentiates the alleles. This figure shows how ASM can be used to predict ASE behavior. Additional haplotyping can associate the SNPs thereby allele specifically linking the methylation status of a regulatory region to its expression. c) An instance of a SNP overlapping with a CpG site in an ASM region. The SNP in this example disrupts a CpG site such that the CpG site no longer exists in the orange allele. The ASM behavior seen in this region may be caused by the SNP itself or the SNP/CpG overlap. d) An instance of a SNP overlapping with a CpG site in a non-ASM region. Similar to c), the overlapped CpG site no longer exists in the orange allele.

1.4.1 Chromosome X Inactivation

The extensive research into chromosome X inactivation (XCI) has illuminated a pathway for allele specific methylation via non-coding RNAs (ncRNAs). To maintain proper dosage specificity in female cells, one chromosome X copy is inactivated upon differentiation. The inactivated chromosome is heavily methylated, compacted into a heterochromatic state, and is largely transcriptionally silent. Experiments have shown that this process is directed by large noncoding-RNAs, mainly Xist and TsiX, which are regulated at the X-inactivation center(*52, 57-62*). Prior to XCI, the two X chromosomes pair at and around the X-inactivation center(*63-65*). The pairing and XCI initiation relies on the binding of Oct4 and CTCF at the XCI region(*63, 66, 67*).

Changes in the allele-specific transcription of *TsiX* and *Xist* loci initiate a cascade that results in the methylation-mediated silencing of one copy of chromosome X via a yet poorly understood counting mechanism(*63, 68, 69*). Transcription of *Xist* is up regulated while *TsiX* is down regulated on the selected X_i chromosome (inactive chromosome X) while on X_a (active chromosome X), *Xist* is down regulated and *TsiX* is up regulated. On X_a TsiX has been shown to associate with DNMT3A at the *Xist* promoter region. This leads to the de novo methylation of the *Xist* promoter and thusly the silencing of *Xist* transcription on X_a(*70*). Xist contains a RepA element to which PRC2 binds and this interaction recruits PRC2 to chromosome X(*71, 72*). As *Xist*

transcription remains high on X_i, Xist localizes to various locations across X_i and its recruitment of PRC2 leads to a chromosome-wide H3K27me3 modification of nucleosomes on X_i(*62, 73*). PRC2 then likely recruits DNA methyltransferases to X_i where it is globally methylated(*74*). The XCI process reveals an RNA directed DNA methylation (RdDM) pathway in human. Experiments have observed active RdDM pathways in Arabidopsis but there are few such mammalian examples(*2, 75-78*).

1.4.2 Imprinted Genes

Imprint control elements (ICE) regulate the allele specific expression of imprinted genes via its own methylated state. Imprinting patterns (i.e. methylation states of ICEs) are formed during gametogenesis with paternal ICEs methylated in sperm and maternal ICEs methylated during oocyte growth(*79-81*). DNMT3A and DNMT3L are required for the establishment of these imprinting patterns in both cell types and DNMT1 is necessary for maintenance of these imprints(*82, 83*). Histone modifications may play a role in the establishment of these methylation patterns. DNMT3A was shown to recognize H4R2me2 modifications via its PHD motif; the recruitment of DNMT3A leads to de novo methylation of nearby sequences, which was shown in the beta-globulin locus(*84*). Another experiment has shown that DNMT3L binds to H3 but it is repelled by the H3K4me3 modification(*85*). This repulsion would prevent de novo methylation of regions marked for poised or active transcription(*86, 87*). The methylation of ICEs results in the transcription

of protein-coding genes in cis (e.g. *GNAS*, *IGF2*, *PWS/AS*, *KCNQ1*). Since DNA methylation is associated with transcriptional silencing, this is an unexpected consequence. DNA methylation at ICEs serves to repress transcriptionally repressive factors, which explains its appearance as an activator. Unmethylated ICEs can form insulators, which prevent mRNA transcription, or unmethylated ICEs lead to the transcription of ncRNAs, which then repress transcription of nearby protein coding genes in a cis fashion(*88-90*).

Imprinting defects, which are seen as the improper establishment of regional DNA methylation haplotypes (i.e. epi-haplotypes), are associated with severe disorders. For example, imprinting defects in chromosome region 15q11.2-q13 are directly linked to Angelman Syndrome (severe developmental delay) and Prader-Willi Syndrome (severe hypotonia)(*91*). Imprinting defects within the imprinted *GNAS1* gene lead to Pseudohypoparathyroidism, which causes electrolyte imbalances and low-levels of calcium in the blood(*92, 93*). Loss of imprinting at the *H19*|*IGF2* region leads to Beckwith–Wiedemann syndrome (large body size, embryonal tumors, and visceromegaly)(*94*). Allele-specific DNA methylation aberrations also extend beyond imprinted genes. Experiments have shown the expression of the X-linked gene *MeCP2* is tightly regulated in vivo. Mutations in this gene cause Rett Syndrome while a 2x over expression of the gene leads to neurological disorders(*95*). Severe phenotypes result from deviations from

expected ASM patterns, which demonstrate the importance of a tightly regulated methylation regulatory network.

1.4.3 General Allele-Specific Methylation

Beyond imprinted genes whose methylation patterns are determined during gametogenesis, there are additional ASM patterns in mouse and human. A human methylation-sensitive SNP analysis involving 12 samples covering 6 tissues revealed 16 ASM candidates(*56*). The authors found that 12 of those sites strongly associated with nearby SNPs. Unlike in imprinted genes where methylation patterns are determined by parent-of-origin, these ASM sites were dependent on sequence. The authors labeled these ASM regions as epi-haplotypes. Further studying ASM in human, a recent targeted genome survey of CpG islands in chromosome 12 and chromosome 20 across 16 cell lines found that many CpGs with methylation frequencies between .25 and .75 (“fuzzily methylated regions”)(*38, 39*). Using SNPs to separate alleles in order to explore areas of fuzzy methylation, the authors found between 23% and 37% of heterozygous SNPs showed significant ASM behavior. Due to the high CpG content of the targeted regions, many SNPs overlapped with CpG sites and many of the ASM examples were due to sequence differences between the alleles (Figures 2c and 2d). For example, a G/A heterozygous SNP at the second position of a CpG dinucleotide would appear as a CG sequence on one allele and a CA on another allele. The CA sequence would then be completely unmethylated based solely on its sequence. Additional

examples of ASM without SNPs at a CpGs were also found. Interestingly, different ASM scales were detected as some regions contained only a single CpG with ASM while other regions showed ASM spanning several CpGs. A comparison across cell lines revealed that only 6% of ASM behavior was conserved across cells that contained the same heterozygous SNP. Additional experiments in mouse have found hundreds of candidate ASM regions genome-wide(*96*). Examining 18 of those sites in detail, the authors found 15 of them showed the same methylation levels in the male germ line and were only differentially methylated in somatic cells. Unlike imprinting, the methylation states of these 15 sites are not determined until after fertilization, which means a mechanism separate than imprinting is available for establishing ASM. Many of the ASM regions in this study were associated with sequence variants, which demonstrate the likely role of sequence specific factors in regulating ASM in mouse(*97-99*). Another genome-wide analysis showed that 2,704 SNPs displayed ASM and 90.3% of them were associated with cis-acting ASM(*100*). ASM flipping has also been observed. ASM flipping describes a SNP region where allele A is more methylated than allele B in one cell line and allele B is more methylated than allele A in another cell line. Instances of ASM flipping were found across cell line families in our targeted bisulfite sequencing study(*39*). These studies validate the presence of cis-dependent ASM in various cell lines in both human and mouse. In the mammalian genome ASM occurs outside of imprinted regions, at different scales, and may be regulated by cis-sequences.

Recent experiments have shown that cell line clonality can lead to biological artifacts and diminishes the *in vivo* applicability of results based on clonal cell lines. A study found that about 20% of tested lymphoblastoid cell lines were pauciclonal (consisting of only a few clones) or monoclonal(*101*). In cell lines where only a few clones are present, sensitive allele-specific studies may not reflect the whole cell population *in vivo*. Such clonal cell lines are likely to exhibit artifacts like random mono-allelic expression, which mask the biological signal found in the *in vivo* cell population. Affected studies include gene expression and methylation studies. Another study found that many traits (e.g. RNA transcript levels, drug responses) are highly variable across lymphoblastoid cell lines and that gene expression is better explained by artifacts (e.g. EBV copy numbers and growth rate) than by genetic variants(*102*). However, another group that investigated multiple individual-specific cell types showed that clonality did not significantly affect their ASE results(*103*). Given these findings, experimentalists should account for the clonality of their investigated cell lines when evaluating results, especially in allele specific studies.

Linked with ASM in many cases, ASE has been observed to occur randomly in clonal cell lines. Experiments comparing the ASE behavior of clonal cell lines derived from B-lymphoblastoid cells revealed the presence of a class of genes whose ASE behavior were random(*104*). This class describes genes that are expressed on the paternal, maternal, or both alleles across clones. Such genes are spread throughout the genome and do not cluster

together in enriched locations. Many of the genes (80%) that showed monoallelic expression in at least one clonal cell line showed biallelic expression in other clonal lines. This demonstrates that these genes can be expressed from either allele or both simultaneously. Overall 5%-10% of human autosomal genes were found to be randomly monoallelically expressed. Additional studies need to be performed to demonstrate the prevalence of random monoallelic expression outside of the B-lymphoblastoid cell lines investigated in this paper. Connecting DNA methylation to ASE remains an interesting challenge but methylation studies have the potential to explain many of the recently observed ASE phenomena, including random monoallelic expression.

1.5 Methylation Pattern Erasure: iPS and PGC Dynamics

Reprogramming involves the activation of a small set of transcription factors within a differentiated cell, which transforms the cell into an ES-like state(*105-108*). The ability to reprogram differentiated cells into a pluripotent state raises hopes for improved transplantation and disease therapies. An issue with transforming cells into induced pluripotent stem (iPS) cell lines is the low transformation efficiency (< .1%). The activation of certain transcription factors during transformation may bias the iPS cell towards certain differentiated cell types or lead to tumorigenesis(*106, 109*). In order for iPS-

based therapies to become widespread, iPS transformation efficiency and safety must be improved(*108, 110-115*).

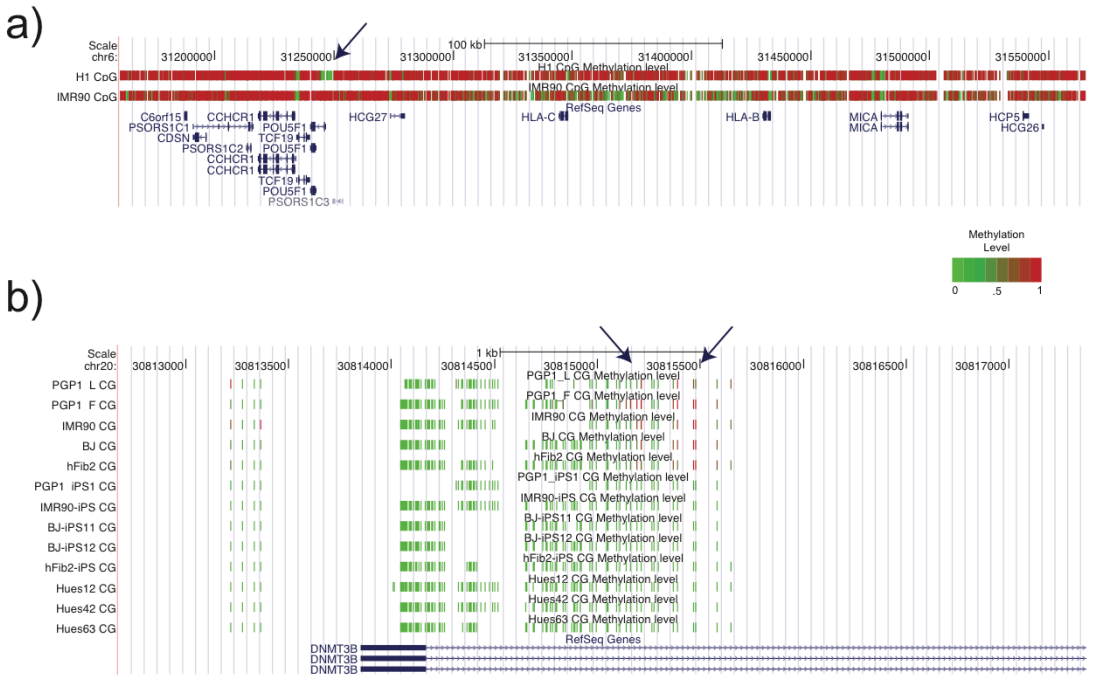


Figure 1.5.1 a) This figure shows the methylation status of H1 (ES cell) and IMR90 (fibroblast cell) across a 3 kb region of chromosome 6. Each vertical bar represents a CpG site whose color is dependent on its methylation level. The arrow indicates the promoter of *POU5F1*, which codes for the Oct4 protein, a master regulator of pluripotency in ES cells. The promoter is unmethylated in H1, where Oct4 is transcribed, and methylated in IMR90, where Oct4 is repressed. Regions upstream of the *POU5F1* promoter show larges swaths of differential methylation. Successfully reprogramming involving large methylation changes in such areas. b) A detailed view of the *DNMT3B* TSS site across 13 cell lines. The arrows indicate CpG sites that show cell type specific methylation. The iPS lines have methylation signatures that match the ES cells instead of their pre-transformed cell types. PGP_1 is a lymphoblastoid cell line while PGP1_F, IMR90, BJ, and hFib2 are fibroblast cell lines. PGP1_iPS1 was transformed from the PGP1 fibroblast. HUES12, HUES42 and HUES63 are embryonic stem cell lines.

1.5.1 Methylation and iPS Reprogramming

During differentiation, which is reversed during iPS transformation, massive changes in methylation occur genome-wide. Genome-wide methylome constructions of a fibroblast, IMR90, and an ES cell, H1, have revealed that many regions of the IMR90 genome are methylated at much lower frequency than in H1(Figure 1.3.1a)(*1*). However, the authors found 491 regions where IMR90 was more methylated than in H1 (Differentially Methylated Regions, DMRs). The 139 genes associated with these DMRs showed higher expression in H1 relative to IMR90 and 113 genes showed lower expression. The majority of these genes had DMRs within 2 kb upstream from the TSS or 5'-UTR. The H1 cell line also showed non-CpG methylation (CHH and CHG) throughout the genome while only CpG methylation was present in IMR90. Relating the methylation signatures of iPS cells to fibroblasts and ES cells, a targeted bisulfite study showed that the methylation signatures of human reprogrammed cells from two fibroblast cell lines clustered closely to ES cells and were distinct from their respective untransformed cell lines (Figure 1.3.1b)(*38*). However, the iPS cell lines were still distinguishable from ES cell lines, showing that differences between the iPS cell lines and ES cell lines exist. This difference was also found in another study where iPS cells had regions with methylation signatures that were dissimilar to both ES cells and their respective untransformed cell lines(*41*). These studies show that there are significant methylation differences between

ES and differentiated cells. Successful iPS transformation involves massive methylation changes in order for a differentiated cell to arrive at a pluripotent state.

1.5.2 DNA Methylation Erasure During Reprogramming

Reprogramming involves DNA demethylation and inhibiting DNA demethylation thusly diminishes reprogramming efficiency. A mammalian protein associated with demethylation is activation-induced cytidine deaminase, AID, which is a 5-meC deaminase. AID was first discovered as a necessary component in Class Switch Recombination and high levels of Somatic Hypermutation processes, which take place within B-cells(*116-118*). Initially thought to be expressed only in immune cells, further research showed that AID was also expressed in pluripotent cells(*119*). AID's association with demethylation stems from its conversion of methylated cytosines to thymines(*120, 121*). The converted base is removed via G:T mismatch repair, which, for example, can be performed by Mbd4(*119, 122, 123*). Demonstrating that the protein AID is important in iPS reprogramming, a group used mouse ES cells fused with human fibroblasts cells to show that AID regulates the transcription of Oct4 and Nanog(*124*). The fusion of a mouse ES cell and a human fibroblast cell creates a heterokaryon and this process efficiently produces iPS cells. 70% of heterokaryons expressed high levels of human Oct4, Nanog, and GAPDH (control) transcripts on the third day after fusion. Noticing that AID transcripts were found in the heterokaryons and that AID

was bound to the methylated promoter regions of Oct4 and Nanog in human fibroblasts, the authors used several siRNAs that recognized different regions on the AID transcript to knock down AID in both the mouse ES and human fibroblast cells. These siRNAs were then transfused into both cells lines 24 hours before the cells were fused. Oct4 and Nanog expression levels were reduced by at least 80% in the AID-knockout heterokaryons relative to the control and bisulfite sequencing revealed significant methylation at the promoter regions of Oct4 and Nanog relative to the control. The results show that the presence of AID likely leads to demethylation at Oct and Nanog promoters, which results in their transcription. Although the knockdown of AID resulted in a significant inhibition of iPS reprogramming, over expression of AID did not change the efficiency of iPS reprogramming. The discovery of AID's activity during reprogramming does not exclude the possibility of other factors being involved in DNA demethylation. Additionally, the details of AID's demethylation mechanism in pluripotent cells are still unknown. Additional research is needed to eliminate these ambiguities and undoubtedly show that AID is a necessary factor in active DNA demethylation. Given these caveats, this experiment demonstrates the role of AID in demethylation is important for the reprogramming of somatic cells into a pluripotent state.

1.5.3 Primordial Germ Cell Methylation Erasure

In additional to affecting reprogramming efficiency, AID has been shown to play a role in DNA demethylation in primordial germ cells (PGCs).

The erasure of DNA methylation patterns, including those in imprinted regions and the inactivated X_i, is important for PGCs as it eliminates epimutations and allows a return to pluripotency(*125-128*). Studies have shown that methylation signatures in imprinted regions, single copy genes, and certain repeats (e.g. LINE1) are significantly demethylated in PGCs between 11 and 13.5 days post coitum(*129*). A recent study expanded on these past findings and looked at methylation effects caused by AID in various genetic elements within murine PGCs(*130*). Demethylation in PGCs was found be global and included gene regions, transposons, and repeats; the final epigenetic state of PGCs at 13.5 days post coitum was termed an “epigenetic ground state”, where the genome is mostly demethylated and histone marks are mostly absent(*126, 130-133*). Comparing $Aid^{-/-}$ knockout to wild-type mice, the authors found that the AID deficient PGCs were significantly more methylated; there was a sex bias as the female AID knockout cell lines were more methylated than the male knockout cell lines. Interestingly the promoters between the AID knockout and wild-type mice did not show a significant difference; transposons, introns, and exons were more methylated in the knockout line.

1.6 Relationships between Transcription Factors and DNA Methylation

Although genome-wide methylomes have recently become available for various human cell lines, the reasons why genomic regions are unmethylated

or methylated are mostly unclear. For example, the observation that CpG islands in the human genome are mostly unmethylated has not been explained on a genome-wide scale. However, studies are beginning to reveal a deeper relationship between specific transcription factors and local DNA methylation patterns (Table 1). These studies serve to build a foundation that will thoroughly explain the non-random nature of DNA methylation. A set of three transcription factors that tightly associate with DNA methylation are discussed below.

Table 1.6.1 Summary of reviewed DNA methylation-related transcription factors.

Transcription Factor	Methylated Binding Specificity	DNA Methylation-related Consequence of Transcription Factor Binding
Cfp1	Unmethylated	(1) Binds to unmethylated CpG islands and (2) recruits Set1a and Set1b complexes, which trimethylates histones at H3K4 independently of RNA PolII binding(*134-136*).
CTCF	Unmethylated	(1) Blocks enhancer-promoter interactions by acting as an insulator(*137, 138*) and (2) protects genomic regions from nearby DNA methylation encroachment(*139-141*).
MeCP2	Methylated	Associates genome-wide with methylated CpGs in mouse. Leads to H3 deacetylation, which results in global chromatin structure changes(*95*). Loss of MeCP2 leads to up-regulation of 2,184 genes and down-regulation of 377 genes in mouse(*142*).
Sp1	Unmethylated	Protects genomic regions from nearby DNA methylation encroachment(*143, 144*).

1.6.1 CTCF

The CCCTC-binding factor (CTCF) contains an 11 zinc finger DNA-binding domain that is highly conserved in higher eukaryotes(*145, 146*). Although initially reported as a silencer, CTCF is a versatile transcription factor that has been seen to enhance and repress transcription at promoters as well as act as an insulator(*137, 147-150*). It binds to a range of sequences and its zinc fingers allosterically customize according to DNA sequence and nearby cofactors, which confers CTCF's ability to recognize a wide variety of sequences(*145*).

CTCF plays a critical role in the establishment and maintenance of the imprinted *H19*|*IGF2* region. Directly upstream of the *H19* locus is an imprint control element (ICE) that contains four CTCF binding sites. This ICE is necessary for the allele specific expression of both *H19* and *IGF2* genes. Methylation of this ICE on the paternal allele prevents CTCF binding within the ICE and also suppresses *H19* expression via promoter methylation(*151-153*). CTCF binding on the maternal allele prevents enhancers downstream of *H19* from interacting with the promoter region of *IGF2*. Without CTCF's insulating function on the paternal allele, enhancers downstream of *H19* are able to interact with the *IGF2* promoter, which leads to the paternal expression of *IGF2*. Hypomethylation of the *H19* promoter on the maternal allele leads to maternal *H19* expression. Recent data has shown that CTCF is an essential factor in the formation of long-range contacts (i.e. loops) between this

imprinted locus and distal enhancer regions(*138, 154*). In addition to mediating long-range regulatory interactions, CTCF can protect nearby regions from encroaching DNA methylation. An experiment looking at the retinoblastoma tumor suppressor (*Rb*) gene found that CTCF binding prevents DNA methylation from spreading into the *Rb*'s CpG island promoter region(*155*). The ability of CTCF to protect promoters from repressive DNA methylation was also seen in the *c-MYC* promoter(*140*) and in the *BRCA1* promoter(*141*). These examples show that CTCF binding depends on the methylation state of its binding sequence and its binding can mediate long range regulatory interactions and affect local methylation patterns.

1.6.2 Cfp1

CXXC finger protein 1 (Cfp1) contains a conserved CXXC domain that is sufficient for specific binding to unmethylated CpG dinucleotides and two plant homeodomain (PHD) fingers(*156, 157*). The yeast analog of Cfp1, Spp1, has been found to bind to the H3K4me3 histone modification, which is associated with poised or active transcription, via PHD finger interactions(*158*). In addition to recognizing H3K4me3 modifications in yeast, Cfp1 associates with the mammalian Set1a and Set1b methyltransferase complexes(*135, 136*), which are known to trimethylate H3K4.

Using ChIP-Seq technology on mouse brain nuclei, a recent study found that Cfp1 was localized at unmethylated CpG islands(*134*). CpG islands bound by Cfp1 were found to also exhibit H3K4me3 modifications.

Unmethylated CpG islands not occupied by Cfp1 or H3K4me3 where found to align with the repressive histone modification mark H3K27me3(*159*). To test the direct link between Cfp1 and H3K4me3 modification, the authors inserted promoterless CpG-rich DNA into regions not associated with H3K4me3. The results showed Cfp1 binding peaks at the DNA sequence insertion sites and H3K4me3 was present as these locations. However, there was no sign of RNA Pol II binding. This test showed that the H3K4me3 modification is not a byproduct of transcription since it occurs independently of RNA Pol II binding. Cfp1 is a transcription factor that binds to unmethylated CpG islands and modifies nearby histones in a manner that promotes RNA Pol II recruitment.

1.6.3 MeCP2

While the previous two transcription factors bind to unmethylated sequences, MeCP2 recognizes CpG methylated DNA(*160, 161*). MeCP2 is apart of the MBD protein family (MeCP2, MBD1, MBD2, and MBD4), which consists of proteins that share a homologous methylated DNA recognition domain but contain different transcriptional repression domains(*162-164*). The mammalian form of MBD3 is an exception since there is a mutation in its MBD that prevents it from binding to methylated CpGs(*165*). MeCP2 binding was originally thought to cause transcriptional repression due to its recruitment of histone modifying enzymes, which results in repressive histone marks around regions bound by MeCP2, namely histone deacetylation and trimethylation of H3K9(*166-168*). MeCP2 is expressed highly only in neurons and its

expression level is tightly regulated. Embryonic *MeCP2*-null mice die at week 12 while mature mice under- or over-expressing *MeCP2* (e.g. heterozygous females) exhibit neurological disorders(*169-172*).

The transcriptional role of MeCP2 has recently changed from a transcriptional repressor to a global chromatin remodeler that can lead to the expression and repression of various genes. Since MeCP2 binding leads to traditionally associated transcriptionally repressive histone modification patterns, MeCP2 binding near genes was thought to silence them. However, a recent study showed that mice over-expressing *MeCP2* led to a significant number of up-regulated genes relative to *MeCP2*-null mice (2,184 genes up-regulated while 377 genes down-regulated)(*142*). These authors found that a global transcriptional activator CREB1 and MeCP2 co-occupied promoters of many activated genes, including the *Creb1* promoter, and CREB1 was also found to copurify with MeCP2. These findings are inconsistent with the model of *MeCP2* as a transcriptional repressor. Further changing the view of *MeCP2's* role as a repressor, another study performed on mouse brain tissue revealed MeCP2 binding correlates with DNA methylation levels genome-wide. MeCP2 is a protein highly expressed in neuronal cells and its concentration in neurons is similar to that of nucleosomes(*95*). Using ChIP-Seq to enrich for sequences bound to MeCP2, the authors found that 56% of the mouse genome showed some MeCP2 binding activity. The intensity of the MeCP2 signal increased as DNA methylation density increased. MeCP2 binding was not found at unmethylated CpG islands, suggesting a genome-

wide methylated DNA binding preference for MeCP2. To test the relationship between H3 acetylation and MeCP2 abundance, the level of H3 acetylation was measured between wild-type and *MeCP2*-null neurons. There was a 2.6 fold increase in H3 acetylation in the *MeCP2*-null neurons relative to the wild-type. As a control, wild-type and *MeCP2*-null glial cells were also examined but no significant differences in H3 acetylation were found. The *MeCP2*-null brain cell lines also showed a 1.6 fold increase in transcription of repetitive elements while no expression change was seen in Actb, c-Myc, or tyrosine hydroxylase genes. These authors present MeCP2 as a protein that binds to methylated DNA on a global scale. In accordance with previous experiments, MeCP2 does reduce H3 acetylation levels and inhibits transcription of repetitive elements. Due to the global scale of histone deacetylation in the presence of MeCP2, MeCP2's association with deacetylation, which is traditionally associated with silencing transcription, and activation of gene transcription are not necessarily mutually exclusive. The consequences of global histone modification changes are unknown and a rigorous study into this subject may bridge these two seemingly opposing characteristics of MeCP2.

1.6.4 Transcription Factors Regulate DNA Methylation and Are Regulated by DNA Methylation

Experiments focused on specific transcription factors have shown the dependence of DNA methylation states on transcription factor binding. However, the list of transcription factors affected by methylation is short and the relationships between DNA methylation and most transcription factors are still unclear. A recent motif based study looked at CpG island sequences that are resistant to de novo methylation in colorectal and leukemia cells. Using motif analysis tools, the authors found a set of motifs that are strongly associated with de novo methylation resistance(*173*). The most significant motifs are YY1, Sp1, and NRF1. Sp1 has been known to protect CpG islands from methylation(*143, 144*). YY1 recruits the Polycomb Repressive Complex 2, PRC2, which is known for transcriptional silencing via the H3K27me3 histone modification(*174*). PRC2 may also act in a context-specific manner as an allele-specific antagonist to DNA methylation(*175*). Elucidating the causes and consequences of DNA methylation will yield a better understanding of cell differentiation, reprogramming to pluripotency, and cancer development since these processes involve massive genome-wide methylation changes.

1.7 Hydroxymethylcytosine: A new cytosine modification

Complementing AID's demethylation activities via deamination and G:T mismatch repair, a new family of mammalian proteins have been reported to convert methylated cytosine (5meC) bases into hydroxymethylcytosine

(5hmC) in vivo (Figure 1.5.1)(*176, 177*). The first study identified an unknown base that started appearing on Thin Layer Chromatography plates(*176*). The authors found that 5hmC co-migrated with this unknown spot and mass spectrometry later confirmed the presence of this compound as 5hmC. This new cytosine modification was found in the nuclei of Purkinje neurons and granule cells. Hmc5 was found to make up .59% of all bases in Purkinje DNA and .23% in granule cell nuclei. A seemingly insignificant fraction of DNA exhibits this modification but relative to CpG, which makes up about 1% of all bases, the abundance of 5hmC is indeed significant. Another study searched for proteins responsible for this novel modified base. Trypanosome proteins JPB1 and JPB2 are known to hydroxylate and glucosylate the methyl group in a thymine, which results in b-D-glucosyl hydroxymethyluracil(*178*). Using the predicted oxygenase domain in those proteins to find proteins with similar functions in human, this study found that the predicted trypanosome oxygenase domain shared significant homology with the human proteins TET1, TET2, TET3. To test for TET1's activity, the authors transfected hemagglutinin-tagged TET1 into embryonic kidney cells (HEK 293). Kidney cells that showed an increasingly strong signal for hemagglutinin (HA) also showed a decreasing signal for methylated cytosine (5meC) relative to a mock control. Using methylation sensitive restriction enzyme techniques, the authors discovered a base of unknown identity. Using mass spectrometry and comparing with the fragmentation pattern of 5hmC, the authors found that the unknown base was hmC. Furthermore, the authors found that 5hmC made up

4 to 6% of all cytosine species at MspI cleavage sites in mouse ES cells while 5meC made up 55 to 60%. Knockdown of TET1 via RNAi led to a 40% decrease in 5hmC levels. The continued presence of 5hmC in the absence of TET1 was attributed to the presence of additional proteins, such as TET2, and TET3. Finally, the induced differentiation of mouse ES cells by removal of LIF for 5 days led to a 80% TET1 transcript decline and about a 40% drop in 5hmC levels. Protocols for large-scale identification of 5hmC have yet been presented but a recent study evaluated the ability of current methylation detection techniques to detect 5meC, 5hmC, or both. 5hmC appears to be resistant to bisulfite and bisulfite sequencing results thusly do not distinguish between 5meC and 5hmC. PCR appears to amplify 5hmC and 5meC sequences with similar efficiency. The monoclonal antibody used in MeDIP experiments(*179-181*), however, is 5meC specific. The proteins MBD2b, MBD1, MBD4, and MeCP2 also bind 5meC specifically(*162, 182-184*). These studies show the presence of another type of modified cytosine in vivo. Bisulfite sequencing cannot discern between the cytosine types but MeDIP and certain protein complexes can. Recent evidence showed that single molecule real time sequencing technology discriminates between the two types of cytosine modifications, too. The presence of 5meC and 5hmC affects polymerase kinetics in SMRT and this can be exploited to differentiate between the two modifications at a base pair resolution(*185*).

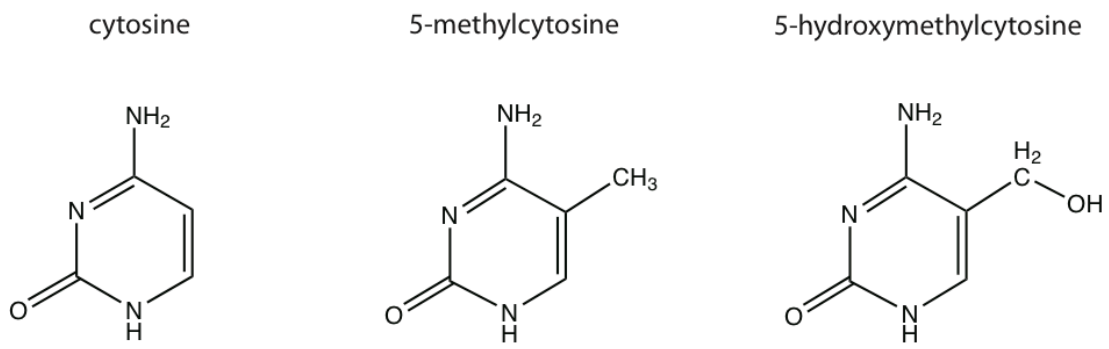


Figure 1.7.1 The structure of cytosine and its two modified forms found in the human genome.

1.8 Conclusion

Methylation is present in various higher eukaryotes but methylation targeting of genomic regions is variable. Vertebrates tend to have highly methylated genomes. In humans, embryonic stem cells tend to be more methylated than differentiated cell lines and human ES cells also exhibit non-CpG methylation (CHH and CHG). Methylation signatures are cell type specific and the methylation frequencies of CpGs can identify human tissues, ES cells, and iPS cells. Methylation signatures can also discriminate between normal and cancer cells and even detect cancer subtypes. DNA methylation can also be allele specific, which differentiates the parental alleles. Chromosome X inactivation and imprinted genes regions are examples of this behavior. Recent research has found that allele specific methylation (ASM) is also found throughout autosomal regions outside of imprinted gene regions. The number of identified allele specifically methylated sites (ASM) is expanding rapidly as genome-wide experiments show that ASM is prevalent throughout the genome.

One reason for the cell type specificity of DNA methylation is due to its interactions with transcription factors. Transcription factors like Cfp1, Sp1, and CTCF bind to unmethylated sequences while other transcription factors like MeCP2 recognize methylated sequences only. Cfp1 binding in CpG islands leads to H3K4me3 modification of nearby histones, which is an RNA PolII positive mark. Sp1 and CTCF can protect certain regions (e.g. promoter

regions) from nearby DNA methylation encroachment. MeCP2 binds to methylated sequences genome-wide and regulates chromatin structure globally. Motif analyses have revealed transcription factor motifs, YY1, Sp1, and NRF, are associated with a genomic region's resistance to methylation. Although the mammalian DNMT protein family has been established as the source of de novo methylation and methylation maintenance, only recently have mammalian proteins been identified to be involved in methylation erasure. AID is one such candidate protein whose presence is important for successful iPS reprogramming and the establishment of an "epigenetic ground state" in primordial germ cells.

DNA methylation investigations have revealed the presence of another modified cytosine type in human, 5-hydroxymethylcytosine. TET1 protein modulates levels of 5-hydroxymethylcytosine base and TET2 and TET3 are suspected to play a similar role. The currently popular technique of bisulfite sequencing does not discriminate between the two modifications but newer single molecule real time sequencing technology shows promise in identifying both 5-methylcytosine and 5-hydroxymethylcytosine individually.

Researchers are slowly unraveling the mammalian proteins involved in DNA methylation and the complex mammalian regulatory mechanisms of which DNA methylation plays a part, which include histone modifications, non-coding RNAs, and transcription factors. Additional work in this field will elucidate the role of this epigenetic marker in the context of the human transcriptional regulatory network. However, even as DNA methylation is not

completely understood, it stills offers clinical applications and guides researches to biologically active regions in the genome.

Chapter 1, in full, is a reprint of the material as it appears in R. Shoemaker, W. Wang, K. Zhang. DNA methylation and its mediators. *Wiley Interdiscip Rev Syst Biol Med*, Epub (Sep, 2010). The dissertation author was the primary author of this paper.

Chapter 2 Development of a Targeted Padlock Probe Set For Methylation Analysis

2.1 Background

Several experimental methods are available for the detection of methylated cytosines. My research utilizes the targeted bisulfite conversion of DNA, which allows for the unambiguous detection of methylation levels at regions of interest (Figure 2.1.1). However, other experimental methods are available and I would thusly like to briefly introduce the more popular alternative methods and explain why bisulfite conversion is the most appropriate for my studies.

2.1.1 Methylation Assays: Restriction Enzyme Digestion

The ability of a restriction enzyme to recognize a methylated cytosine can be exploited to ascertain the cytosine's methylation state within a restriction enzyme recognition site. There are known pairs of restriction enzymes that share the same recognition sequence but one enzyme is sensitive to methylation (i.e. does not cut at sites containing methylated cytosines) while the other is not. These enzymes are added to separate samples consisting of the same input DNA. The presence of DNA cut by the methylation sensitive enzyme relative to the methylation insensitive enzyme

reveals the methylation level at those enzymes' shared recognition sites. HpaII (methylation sensitive) and MspI (methylation insensitive) are isoschizomers (restriction enzymes with the same recognition sequence) used in these methylation assays. The shared recognition site is CCGG and the two-centermost bases form a CpG site. Samples cut by the MspI enzyme form a control DNA fragmentation distribution since MspI digests irrespective of methylation status. HpaII will cut at unmethylated restriction sites and a sample showing more HpaII fragments thusly reflects more unmethylated recognition sites. HpaII restriction sites are always a subset of the MspI restriction sites. Electrophoresis reveals the general presence of more or less HpaII fragments relative to MspI fragments. If studying a specific site, an experimentalist can use PCR primers that cover CCGG sites in his/her region of interest for a more specific methylation analysis. Microarrays or sequencing can also be used to identify fragments, which allows for a genome-wide survey of methylation via restriction enzyme digestion.

Two applications of this technology are CHARM(*186*) (Comprehensive High-throughput Arrays for Relative Methylation) and HELP(*187, 188*) (HpaII tiny fragment Enrichment by Ligation-mediated PCR) . The HELP method uses the restriction enzymes HpaII and MspI to decipher the methylation status of sites recognized by these isoschizomers. The fragments generated by the restriction enzymes are then ligated with PCR adapters and sequences between 50 bp and 2 kb are amplified. A microarray then reveals the identity of these fragments and the methylation state, as found in the HpaII sample

relative to the MspI sample, can be determined. In the CHARM experimental protocol, the authors use a single restriction enzyme to determine methylation states. CHARM uses the enzyme McrBC, which promiscuously cuts methylated DNA and recognizes the sequence $R^{m}C(N)_{55\text{-}103}R^{m}C$ (^{m}C represents a methylated cytosine). In CHARM, McrBC digested and size selected fragments are hybridized along with a negative control (i.e. size selected DNA not digested with McrBC) in a two-color array. Instead of using a methylation insensitive enzyme as a control, the experiment uses a sample without restriction enzyme digestion for comparison. Like HELP, a microarray that covers CpG regions of interests is used to identify methylated regions.

2.1.2 Methylation Assays: MeDIP and BS-Seq

Me-DIP (Methyl-DNA immunoprecipitation) involves enriching methylated CpG-containing DNA fragments via a ^{m}C specific antibody(*179*). The experimental protocol is similar to chromatin immunoprecipitation except the antibody in Me-DIP targets methylated DNA directly. The Me-DIP enriched DNA is then compared to a negative control (DNA not targeted by antibody) via a microarray or via high-throughput sequencing. Me-DIP requires specific conditions for optimal results. These conditions include using DNA fragments between 300 and 1000 bp long and the availability of single stranded DNA as the antibody preferentially binds to methylated ssDNA (single stranded DNA). Re-annealing is an issue given that methylated cytosines are often present in

high CG content fragments; cytosines and guanines form the strongest network of hydrogen bonds in dsDNA.

Unlike Me-DIP or restriction enzyme digestion, BS-Seq (bisulfite-sequencing) does not rely on relative enrichment calculations but a direct counting of individual cytosines' methylated states. Bisulfite is a chemical that coverts cytosine into uracil via deamination. The bisulfite conversion process greatly prefers unmethylated cytosines to methylated cytosines. During post-bisulfite conversion PCR amplification, the converted cytosines (i.e. unmethylated cytosines) are complemented with adenine while methylated cytosines are complemented with guanine. After amplification, unmethylated cytosines thusly appear as thymines while methylated cytosines appear as cytosines. Similar to Me-DIP, bisulfite conversion prefers ssDNA. Additionally, the timing of the conversion process and the concentration of bisulfite used in the reaction are important. Higher bisulfite concentrations can lead to fragmentation of DNA and longer bisulfite conversion reaction times may lead to the unintended conversion of methylated cytosines. In optimal conditions, bisulfite conversion efficiency is around 98-99%, which is a metric that measures the conversion of cytosines to thymines in sequence contexts that are not known to be methylated(*1, 2, 38, 189, 190*). Bisulfite converted DNA can be sequenced and mapped to a reference genome. Once mapped, the methylation state of the cytosines on the reference genome can be estimated based on the base found in the bisulfite converted (BSC) sequences at the same position. Large-scale experiments have used shotgun BS-Seq to

determine the methylation states of the vast majority of cytosines in a species of interest (*1, 2, 189*).

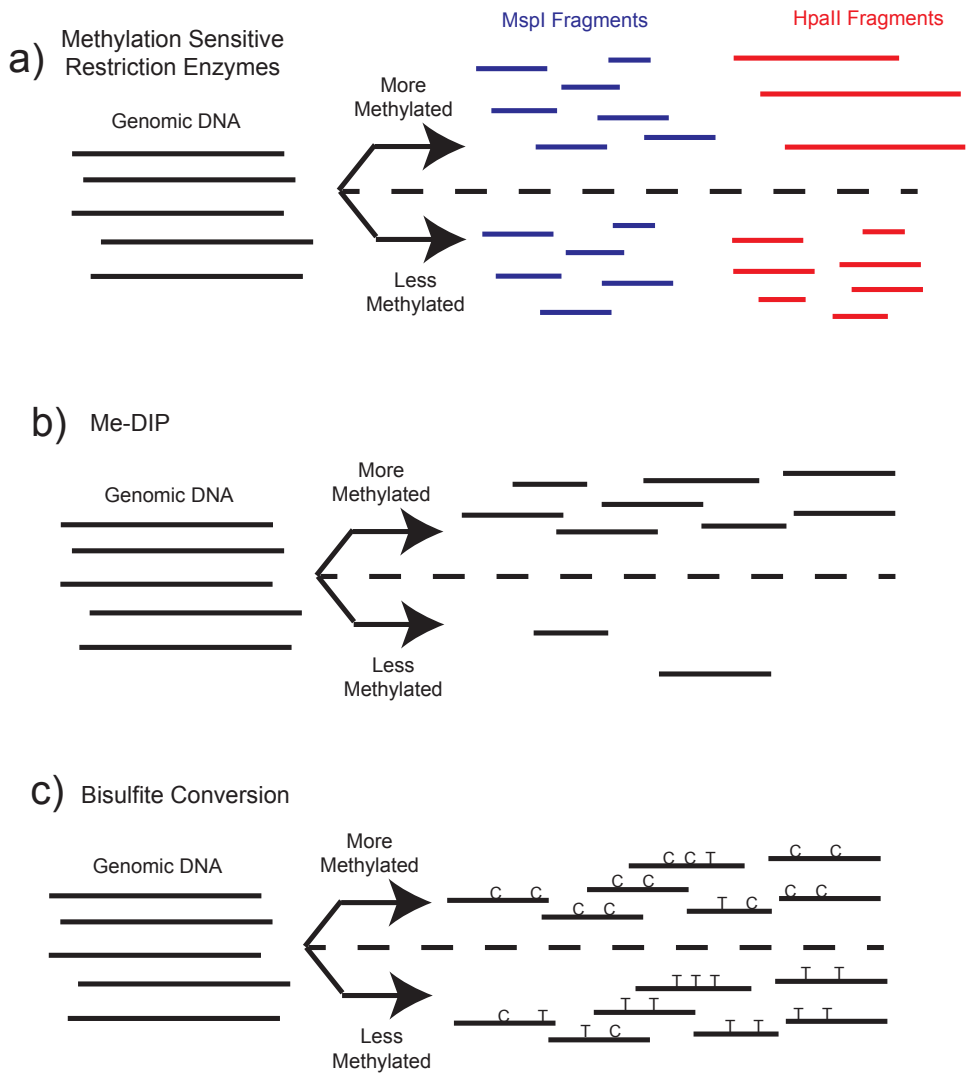


Figure 2.1.1 Each subsection describes the results from an experiment that contains either more methylated or less methylated genomic DNA. A dotted line separates these two experimental conditions. a) Isoschizomeric enzymes fragment DNA differently based on the presence of DNA methylation. MspI (recognizition site CCGG) is methylation insensitive and it thusly produces DNA fragments independent of DNA methylation. HpaII, however, does not cut at methylated sites and thusly HpaII produces less fragments in highly methylated regions. Differences between MspI and HpaII DNA digestion reveal DNA methylation information. b) Me-DIP uses antibodies that recognize methylated cytosines. These antibodies and their attached sequences are then purified from other sequences to which no antibody is attached. More sequences are thusly purified if a strongly methylated region or cell type is studied. The sequences are then compared with a control channel (not shown) that does not use any antibodies. The signal differences between the control and the antibody channels reveal methylation information. c) Bisulfite conversion does not rely on fragmentation or purification. Methylation information is obtained at the sequence level. Since bisulfite converts cytosines to uracils, which appear as thymines after PCR, more methylated sequences will show a higher percentage of cytosines at CpGs than less methylated sequences. In the figure, the cytosine and thymine base are shown at CpG locations only.

2.2 Background of Targeted Bisulfite Sequencing

Targeted bisulfite sequencing allows for the targeting of unique genomic regions so that these regions can be sequenced at a high read depth. Since less sequencing is needed due to the fact that only subset of the complete genome is amplified, sequencing costs are reduced. Since targeted bisulfite sequencing reduces sequencing costs and time, experiments that probe many different cell lines are more easily performed. The ability of this method to target certain regions is due to the use of padlock probes, which are about 250 bp-long nucleotide sequences. Each padlock probe consists of targeting arms, which are 18-30bp long and located at both ends of the padlock probe sequence. Primer sequences and restriction sites connect these two targeting arms together. The targeting arms on a padlock probe are designed to hybridize to sites that are 175-225bp away from each other and the primer sequences allow for PCR amplification of the region located between the targeting arms. In targeted bisulfite sequencing, these arms are designed to hybridize to bisulfite-converted sequences and amplify regions containing CpG sites of interest. Hundreds of thousands of padlock probes can be added to genomic DNA within a single test tube, allowing for the interrogation of many CpG sites of interest (Figure 2.2.1).

The original protocol for targeted bisulfite sequencing consists of 7 general steps: (1) the isolation of genomic DNA, (2) bisulfite conversion of DNA, (3) the addition of padlock probes to BSC DNA, (4) PCR amplification,

(5) removal of padlock probe linker sequences via restriction enzyme digestion, (6) random fragmentation of targeted DNA, and (7) ligation of adaptors for short-read sequencing. Size selection may occur after steps 5, 6, and 7 to eliminate unwanted sequences. The addition of padlock probes after bisulfite conversion complicates padlock generation since unmethylated sequences consist only of A, T, and G bases and potentially methylated sequences (e.g. CpG sites) may consist of A, T, C, and G bases. Since PCR amplification is required immediately after padlock probe addition in order to amplify the targeted regions and non-BSC PCR amplicons do not contain methylation information, BSC must occur before padlock probe addition. When BSC occurs first, unmethylated cytosines are converted to uracil while methylated cytosine remain unchanged. PCR amplicons will reflect the base changes undergone during BSC (i.e. a PCR product will contain a thymine at an unmethylated cytosine site or a cytosine at a methylated cytosine site).

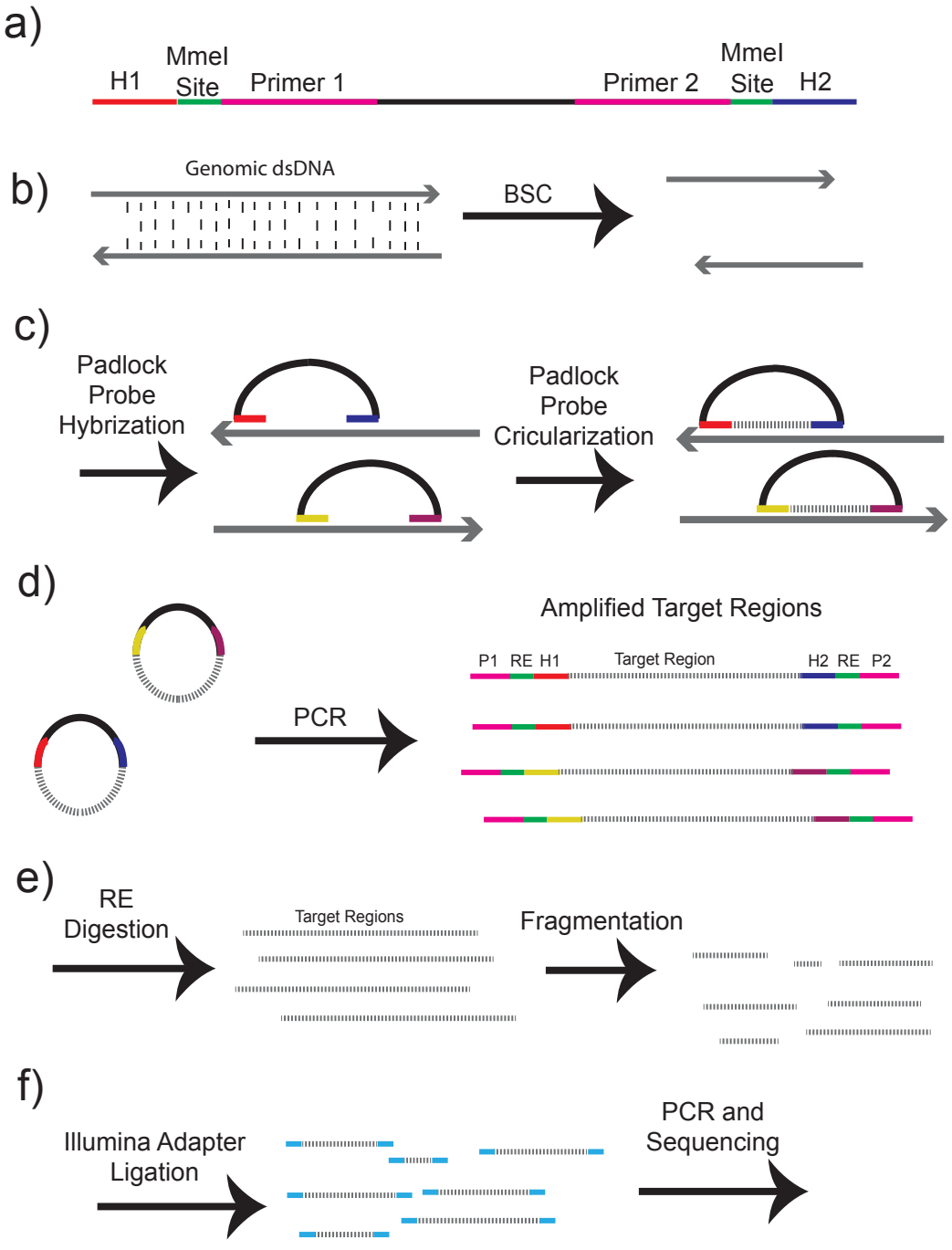
In addition to the target region, the PCR amplicons contain primer sequences and the targeting arm sequences. Adding a class II restriction enzyme site in the padlock probes allows for removal of these uninteresting sequences. This digestion results with longer target region fragments that can be separated from the shorter primer and targeting arms sequences. PAGE gel electrophoresis allows for the separation of these two types of fragments based on their length in a process called size selection. Since the expected size of the targeted regions is known, DNA fragments whose length is close to this expected length are purified from undigested products. Size selection is

important because in the next steps, the DNA is fragmented and ligated to next generation sequencing adaptors. If undesired DNA fragments were not eliminated in the size selection step, they could be ligated to sequencing adaptors and sequenced.

The digestion of size selected DNA fragments allows short read sequencing to access areas within the larger targeted region that it would not otherwise be able to completely sequence. Digestion can occur in various ways, including random shearing or via enzymes. Since target regions span 175-225 bp and short read sequencing can sequence 36-100 bp, areas within the center of many targeted regions cannot be sequenced without fragmentation. Fragmentation will reduce a single target region into several smaller regions. Each of these smaller regions will be sequenced separately. Regions that appear in the center of a long target region may now appear closer to the 5’ and or 3’ end of a fragment, which makes it possible to sequence. These fragments are sequenced using Illumina technology. Once the sequences are known, the BSC sequences are mapped to a reference genome and the methylation states of cytosines are interrogated at base pair resolution.

Figure 2.2.1 a) This figure shows a schematic of a padlock probe nucleotide sequence. The ends are targeting arms, H1 and H2, which are designed to hybridize to the perimeter of 175–225 bp long target regions. These targeting arms are flanked by an MmeI restriction enzyme site. Flanking the MmeI sites are primer sequences. A linker sequence connects the two primer sequences. b) Genomic DNA is initially bisulfite converted. The dotted lines between the dsDNA strands represent base pairing. After BSC, unmethylated cytosines are coverted to uracils, which do not base pair with guanines. The Watson and Crick strands are differentiated by arrows pointing in opposite directions. c) Padlock probes are mixed with the BSC DNA and hybridize to their target regions. Targeting arms are shown in different colors. Targeted regions can overlap if padlock probes target those regions on different strands. 5' to 3' base addition via DNA polymerase followed by ligation via ligase create circularlized padlock probes, which contain their respective target region sequences. Endonucleases are added that disintegrate non-circuralized DNA sequences. d) The circularized probes are PCR amplified. The amplicons contain the primer sites, the MmeI restriction enzyme site (RE), and targeting arms at each end. Size selection is performed after PCR to remove any artifacts. e) The MmeI restriction endonuclease is added to the amplicons, which cuts 18-20 bp away from its recognition site. This digestion removes the primer and targeting arm sequences from the target region sequences. Size selection allows for the purification of targeted regions after digestion. The target regions are then fragmented. This can occur via shearing (i.e. random fragmentation) or via USER enzymes (Uracil-Specific Exicision Reagent), if dUTP is added during the PCR amplification step. USER fragmentation is a semi-random fragmentation technique since sequences are fragmented wherever uracil has replaced thymine in the PCR amplification reaction. USER fragmentation variability is achieved by adding both dUTP and dTTP to the PCR reaction thusly allowing either thymine or uracil addition. Size selection can isolate fragments within a desired bp length window. f) Illumina sequncing adapters are ligated to the ends of the fragments and prepared for sequencing.

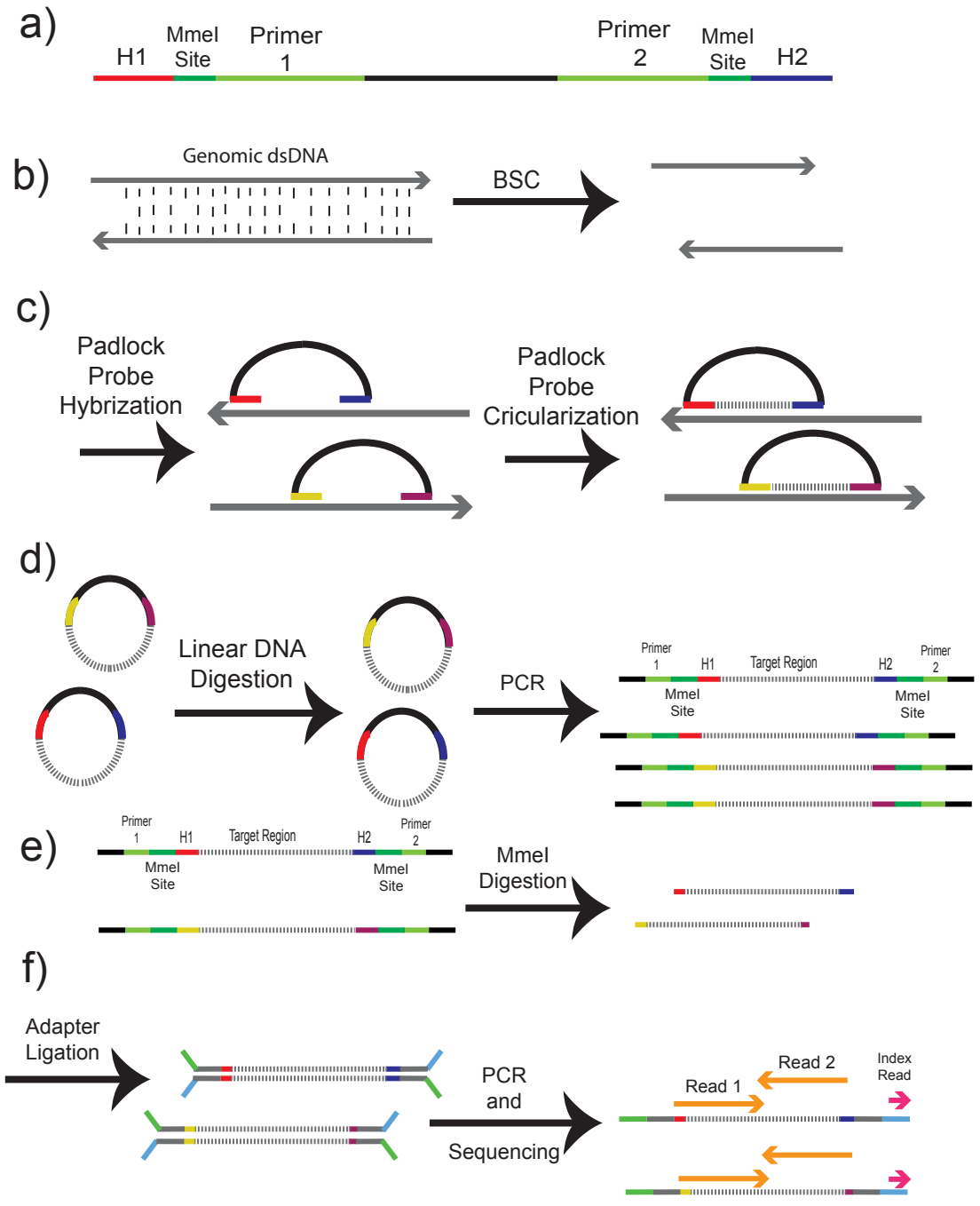
Padlock Probe Overview



A newer targeted bisulfite sequencing padlock probe protocol has been developed that improves the capture experiment (Figure 2.2.2). In the original protocol, the target regions were fragmented and then the target regions were single end sequenced. This new protocol avoids target region fragmentation. Relying on longer read lengths and paired end sequencing, the target regions are only sequenced from the ends. As sequencing lengths reach 100 bp, this means that a target region of 200 bp can be completely sequenced without fragmentation. Looking at the original protocol in Figure 2.2.1, the new protocol modifies steps e) and f). The elimination of the fragmentation step yields a tight distribution of target region lengths. These target regions will appear as a tight band during size selection, allowing for easier extraction from the PAGE gel during size selection. Without fragmentation, each region is represented by 2 reads instead of many reads, which also reduces the number of needed reads per sample. Reads are barcoded in this protocol, which allows for multiple samples to be sequenced in a single flowcell. The new protocol includes recent advancements in next generation sequencing and thusly improves the BSPP capture method.

Figure 2.2.2 a) This figure shows a schematic of a padlock probe nucleotide sequence. The ends are targeting arms, H1 and H2, which are designed to hybridize to the perimeter of 180 bp long target regions. These targeting arms are flanked by MmeI restriction enzyme recognition sites, which are then flanked by common primer sites. A linker sequence connects the two primer sequences. b) Genomic DNA is initially bisulfite converted. The dotted lines between the dsDNA strands represent base pairing. After BSC, unmethylated cytosines are coverted to uracils, which do not base pair with guanines. The Watson and Crick strands are differentiated by arrows pointing in opposite directions. c) Padlock probes are mixed with the BSC DNA and hybridize to their target regions. Targeting arms are shown in different colors. Targeted regions can overlap if padlock probes target those regions on different strands. 5' to 3' base addition followed by ligation creates circularlized padlock probes, which contain their respective target region sequences d) . Endonucleases are added that disintegrate non-circuralized DNA sequences. The remaining circular DNA products are then PCR amplified e) The PCR amplified sequences are digested with MmeI, which removes the primer sites and most of the H1 and H2 targeting arms. f) Y-adapters are ligated to the MmeI digested DNA, which are shown as double stranded molecules in this subfigure. The target regions are then sequenced in a three-step process that entails sequencing read 1, sequencing the index read, and sequencing read 2. The index read contains a unique nucleotide sequence that associates the read with a particular experiment. Multiple experiments can be sequenced in a single flowcell lane with this protocol.

Cpg220k Padlock Probe Overview



2.3 Challenges in Targeted Bisulfite Sequencing

Appropriate selection of targeting regions is a crucial step for the successful application of the targeted bisulfite sequencing method. The targeted regions must allow the experimentalist to test his/her DNA methylation-related hypothesis. For example, an experimentalist may want to test if the methylation states of promoter regions related to DNA damage response genes are changed in response to a particular extracellular stimulus. If the bisulfite targeted padlock probe set does not contain any of these genes' promoter regions, the targeted sequencing approach given this probe set will not help the experimentalist. Currently, generation of a padlock probe set is expensive and time consuming. These constraints encourage padlock probe sets to cover regions that are as broad in biological functionality as possible so that the probe set can be used in many different experiments.

The purpose of targeted bisulfite sequencing is to produce a uniform distribution of sequences from various genome-wide targeted regions from which methylation information can be derived. However, some regions are more efficiently targeted than others. This is partly due to the fact that the human genome is not uniformly accessible and the efficiency of hybridization between the padlock probes and the genome varies as the composition of the bases involved varies. Hybridization may also occur at regions that do not perfectly match the targeting arms' sequences (i.e. non-specific binding). In some cases, if a targeting arm recognizes the majority of a genomic

sequence, it will hybridize to this location even if mismatches are present. Mismatches are bases that are not reverse complementary but are present at the same relative position on the targeting arm and genome. Predicting the ability of a targeting arm to hybridize to a partially matched sequence is difficult.

Bisulfite conversion also introduces additional challenges by reducing the information content of the genome (unmethylated cytosines are converted to thymines). Regions containing cytosines can be represented as a cytosine or a thymine after bisulfite conversion. Sequences that are unique in the human genome (i.e. a sequence is found only at a single location) may no longer be unique after bisulfite conversion. However, the majority of the human genome is methylated(*1*), which means bisulfite conversion in most genomic areas of the genome will not lead to a reduction in base complexity. Targeting areas that are statically methylated will not reveal interesting biology and targeted regions are thusly much more likely to contain mixed methylation states; the base composition at these regions is more likely to change during bisulfite conversion.

If a targeting arm covers a CpG site, two versions of that targeting arm are needed so that the arm can hybridize regardless of the methylation state of that CpG. The targeting arm designed for the unmethylated CpG would contain an adenine at the cytosine position of the CpG since adenine base pairs with the uracil. The targeting arm designed for the methylated CpG would contain a guanine at the cytosine position of the CpG since bisulfite

conversion would not affect methylated cytosines. If both targeting arms each cover a CpG site, 4 padlock probes for this targeted region are produced. If the targeting arms contain a sum of n CpG sites, 2^n probes must be produced to cover all possible CpG methylation combinations. When generating padlock probes, CpG sites in the targeting arms are avoided so that given a certain number of padlock probes, a maximum number of regions are targeted.

If CpG-containing targeting arms are necessary for an experiment (e.g. an experiment that surveys regions with high CpG density, i.e. CpG islands), assumptions of correlated methylation states can be made to reduce the number of padlock probes produced for these regions. This means one assumes that the CpGs on a targeting arm or all methylated or unmethylated based on a correlated methylated assumption. This means if the sum of covered CpGs regions on the targeting arms exceeds 2, only 4 padlock probes need to be produced.

Both targeting arms of a padlock probe must hybridize to the targeted region for amplification of that region to occur. If only one arm hybridizes, the region will not be amplified. For example, the reverse complement sequence of a targeting arm one may be present at numerous chromosomal locations while the reverse complement sequence of the other arm (targeting arm 2) is present only once. In this case, the padlock probe can be tethered to many sites recognized by targeting arm 1 but only the region that is also recognized by targeting arm 2 would be amplified. If too many padlock probes were tethered in regions missing a recognition sequence for targeting arm 2, the

region would be amplified less effectively than regions where the target arms had unique recognition sequences. Given that the human genome can contain more than one copy of a region, padlock probes may target regions that contain identical targeting arm recognition sequences but whose regions may contain sequence variations or are present in different chromosomal locations. These different chromosomal locations may exhibit different epigenetic contexts. For example, one region is in a heterochromatic state while another is in a euchromatic state. Assuming uniform PCR amplification, the methylation information from this targeted region becomes an average across its many genomic instances.

PCR is a necessary evil in the targeting padlock probe protocol. It is necessary so that a sufficient amount of DNA is produced for the size selection and sequencing steps. However, PCR amplifies sequences unevenly and thusly leads to artifacts. For example, region A may be amplified much more efficiently than region B and thus the PCR amplicon population will contain many more copies from region A than region B. Additionally, some regions may contain CpGs that are methylated in only a percentage of the sample cell population. PCR amplification bias may amplify one methylation state of these CpGs more efficiently than the other. The more efficiently amplified methylation state would appear to be dominant. For example, a CpG site may have an actual methylation frequency of .5 (50% of the cytosines at this site are methylated) but PCR preferentially amplifies the unmethylated

strand. The biased amplification would result in a calculated methylation frequency of less of than .5.

2.4 Goals of the Cpg220k Probe Set

The Cpg220k probe set is the first mainstream padlock probe-based DNA methylation assay geared towards commercial production. Agilent, a company that has synthesized past padlock probe sets, agreed to sponsor the production of the Cpg220k probe set so that it can examine the commercial viability of BS padlock probes. Agilent's sponsorship demonstrates confidence in this new method, which was first published by Kun Zhang's lab in 2009(*38, 191*). The two main goals of this new probe set were to expand upon its coverage of regions found to exhibit variable DNA methylation levels across cell lines and to improve its ability to analyze allele-specific methylation.

Offering 3x more coverage than the last generation probe set, the Cpg220k probe set expands CpG coverage and includes newly discovered regions exhibiting differential methylation across various cell types. The first published BSPP (bisulfite sequencing padlock probe set) from Kun Zhang's lab consisted of 30,000 probes and focused mainly on CpG islands on chr12 and chr20 but also covered the transcription start sites for a small number of genes(*38*). This study found that most of the CpGs were consistently unmethylated in ES and fibroblast cell lines; most of the targeted CpGs in this probe set offered no interesting methylation information for fibroblasts and pluripotent-like cell types. Shifting the targeting focus away from CpG islands,

a newer BSPP, Cpg97k, consisted of 97,000 probes and targeted regions demonstrated in outside experiments to exhibit tissue specific methylation patterns.

Experiments using DNaseI have revealed exposed chromosomal locations where DNA-protein interactions are more likely to occur. Additionally, complete methylomes have been produced for a human embryonic and a fibroblast cell line(*1*). Hypersensitive sites are candidate DNA-protein interaction sites and they represent genomic regions that are easily accessible to a digestion enzyme(*192-194*). Since hypersensitive sites reveal sites of likely DNA-protein interactions, the methylation states of these regions are of greater interest. Chapter 1 described several proteins whose recognition of DNA sequences is dependent on those sequences' methylation states. Complimenting the hypersensitivity data, the publication of two human methylomes in late 2009 allows for the direct identification of differentially methylated sites between a pluripotent and a terminally differentiated cell line. The inclusion of new regions derived from hypersensitivity and BSC shotgun sequencing experiments into a larger probe set that contains tissue specific methylation sites presents opportunities for novel DNA methylation discoveries.

The Cpg220k probe set allows not only for improved cell type specific DNA methylation analysis but it also is designed to improve the detection of allele-specific methylation (ASM). ASM defines a set of regions whose methylation states are significantly different, as determined by a statistical test,

between the maternal and paternal chromosomes. In order for ASM analyses to work, one must be able to separate the paternal and maternal alleles. A common way to distinguish between parental alleles is through heterozygous SNPs. Heterozygous SNP sites are sites where the alleles have a different base at the same chromosomal coordinate. Sequencing data with high read depth (usually 10x or higher) can allow for such SNPs to be identified. Additionally, read sequences that contain heterozygous SNPs may also contain CpG sites. These SNP-associated CpG sites can be tested for the presence of ASM. A padlock probe must cover a heterozygous SNP site and at least one CpG site in order for SNP-based ASM analysis to occur. The Cpg220k probe set was thusly designed to include nearby SNP sites so that sequencing reads are more likely to contain both a SNP, if nearby, and at least one CpG site.

A recently developed algorithm, which is an adaption of the r^2 metric from Linkage Disequilibrium analysis, allows for the discovery of organized methylation patterns without SNPs; these organized methylation regions include ASM regions(*39*). In order for this algorithm to work, sequencing reads need to contain at least two CpGs. The presence of more than one CpG on a read allows for the algorithm to calculate the correlation of methylation states between the CpGs on the same sequence. This LD-based algorithm is described in more detail in Chapter 4. The Cpg220k probe set makes this LD algorithm more useful by encouraging the placement of padlock probes that

cover several CpGs within regions that likely exhibit interesting methylation behavior.

The design process of the Cpg220k probe set is an optimization between several factors, which are: (1) the expanded coverage of CpGs found to be differentially methylated and (2) improved CpG-SNP association, (3) the costs of producing a padlock probe set, and (4) the discussed challenges in padlock probe experiments. To target regions that contain SNPs and CpGs of interest, hundreds of thousands of padlock probes must be produced that, for example, contain unique targeting arm sequences and are predicted to hybridize efficiently. Some regions deemed to be of interest for DNA methylation analysis cannot be effectively targeted with padlock probes due to hybridization or sequence degeneracy issues. In such cases, multiple overlapping padlock probes can be produced to overcome the challenges that a single padlock probe would experience. However, the use of additional padlock probes increases the production cost of the probe set. Each padlock probe costs money to synthesize and these costs can skyrocket when producing probe sets that contain of hundreds of thousands of probes. A probe set that minimizes cost and maximizes biologically relevant results is ideal. Ultimately, the Cpg220k probe set is a combination of expanded coverage and improved quality, which I expect will generate the most interesting results yet seen from a targeted bisulfite sequencing experiment.

2.5 Target Region Selection

The regions selected for the Cpg220k probe set were based on several data sets produced within the past two years. I will discuss the data sets used for targeting and why they were selected for this probe set. The main motivation is to select CpGs that show differential methylation across cells exhibiting different phenotypes. The belief is that the methylation patterns affect the transcriptional regulatory network, as described in Chapter 1, and thus the differences in methylation patterns reveal transcriptional regulatory changes. Although the data sets used compare DNA methylation differences across cell lines, we are also interested in looking at ASM, which describes DNA methylation differences within a cell itself.

2.5.1 Tissue Specific CpG Methylation Patterns

In 2009, the Andrew Feinberg group published a series of experiments that explored the methylation levels of CpGs across human liver, spleen, brain, colorectal cancer (and a healthy tissue control), fibroblasts, and iPS transformed fibroblasts(*40, 41*). All CpG islands and non-repetitive lower CpG density regions were examined in these experiments. The authors found 16,379 differentially methylated regions (DMRs) among the three normal tissues, 2,707 DMRs regions in the colorectal cancer samples versus healthy tissue controls, 4,401 DMRs regions in transformed fibroblast cell lines compared to the original fibroblast cell lines, and 71 DMRs between fibroblast

iPS cell lines and human embryonic cell lines. The inclusion of these regions into the Cpg220k data set will allow it to differentiate between cell types based on DNA methylation patterns. Since the Feinberg's group research shows significant overlap between cancer-related DMRs and tissue-specific-DMRs, targeting these regions will improve the probe set's ability to discriminate between cell types that exhibit diseased phenotypes.

2.5.2 Human Methylome Data Sets

Joe Ecker's group published the methylomes of two human cell lines: H1, an embryonic stem cell line, and IMR90, a fibroblast cell line(*1*). This group was the first to publish a complete methylome of a human cell line. Each cell line contained billions of 87bp long reads, which were produced by shotgun BS sequencing. This data allowed for the methylation states of CpGs genome wide to be calculated at base pair resolution. In the previous data sets, the Feinberg lab used a restriction enzyme-based method to detect methylation levels, which is not as precise as bisulfite sequencing. Across these two cell lines, over 6 million CpGs were differentially methylated. Since the methylation comparison was between a pluripotent and a differentiated cell line, these DMRs are assumed to more likely involved in the differentiation or pluripotency process.

2.5.3 Predicted Enhancer Sites (IMR90 and H1)

In additional to the publication of the methylomes of IMR90 and H1, Bing Ren's group published histone modification data on these cell lines. Bing Ren's group demonstrated earlier that certain histone marks, namely H3K4me1 and H3K4me3, are indicative of enhancer and promoter regions, respectively(*87, 195*). Expanding on these findings, another analysis used this IMR90 and H1 histone modification data to predict enhancer regions in both cell lines(*196, 197*). CpGs contained within these predicted enhancer regions were included in the Cpg220k so that they could be studied across various cell lines. Richard Young's lab has recently shown that inter-chromosomal loops are critical for transcriptional regulation in murine cells since they connect distal enhancers to their associated promoter regions(*198*); the transcription of many genes is dependent on the presence of an enhancer-promoter interaction.

2.5.4 Cancer AML Data

Investigating 50,000 CpG sites across 14,000 genes, a recent study demonstrated that DNA methylation patterns can distinguish between sub-types of acute myeloid leukemia and these patterns could be used as a prognostic tool(*42*). This study used 344 newly diagnosed patients' blast cells and sorted these samples into 16 clusters based on DNA methylation patterns. 11 of these clusters contained epigenetic or genetic lesions or were already known but the remaining five clusters could not be associated with any specific features. This data set focused on promoter regions, where DNA methylation

is known to be inversely correlated with protein-coding gene expression. Differentially methylated CpGs in these data sets are thusly likely to be linked to changes in gene expression that are significant in acute myeloid leukemia; these gene expression changes may also be significant in other cancers, too. DM (differentially methylated) CpGs in this data set were included in the CpG220k probe set to improve the probe set's ability to identify unique cancer phenotypes by complimenting the DMRs found in colorectal cancers investigated by Feinberg's group.

2.5.5 DNaseI Hypersensitive Sites and CTCF Binding Sites

John Stamatoyannopoulos' group at the University of Washington provided data for the CpG220k probe set. Identifying sites that are sensitive to the restriction enzyme DNaseI(*199-202*), this group has discovered HSS ([DNaseI] hypersensitive sites) that are cell type-specific and also ones that are conserved across cell types. HSS are regions where DNA is accessible to nucleases; this availability reveals that this region is not protected by nucleosomes or other constitutively bound proteins. These exposed regions are likely enriched in transcription factor binding sites or thought to be involved in other activities, such as chromatin looping(*203, 204*). Since these regions are likely involved in transcriptional regulation, CpG sites within these regions are of greater interest since their methylation states are likely to affect DNA-

protein interactions. Given the wide breadth of DNaseI data from Stamatoyannopoulos' group, these Cpg220k covered regions will likely reflect cell-type specific methylation patterns and promoter ASM for genes that are allele-specifically expressed.

In addition to HSS data, the group also provided CTCF binding sites in B-cells. CTCF, as described in Chapter 1, is DNA binding protein associated with the facilitation of distal regulatory element interactions via chromatin loop formation(*198*). CTCF is a methylation sensitive transcription factor (i.e. CTCF does not binding to methylated sequences) and methylation thusly affects these CTCF-mediated distal interactions (e.g. enhancer-promoter interactions). Although the CTCF binding information is for one cell type only, recent evidence showed that CTCF binding sites are conserved across cell types(*195*). Covering the B-cell CTCF bound regions, the Cpg220k probe set examines the methylation levels at CTCF binding regions; methylation pattern differences in these regions are likely to reflect in different phenotypes due to methylation's effects on CTCF binding.

2.5.6 Imprinted Genes

CpGs within known imprinted genes were also targeted in the Cpg220k probe set. Specific regions, known as imprint control elements, that are within or nearby imprinted genes are expected to be allele specifically methylated in healthy cell types and these regions can thusly serve as a control. Since these known imprinted regions are well studied already, deviations from expected

methylation behavior would be interesting. The discovery of any unexpected behavior may help further improve the mechanistic understanding of imprinted ASM in human.

2.6 Cpg220k Probe Set Generation

I firstly generated a list of all CpG coordinates in the human genome (hg18). I then labeled all CpGs coordinates that belonged to any of the previously mentioned datasets. Using SNP information from the snp130 database, SNPs that were within 70bp of any of the specially labeled CpGs were added to the coordinate list. I used 70 bp as a cutoff since recent Illumina sequencing experiments have been able to produce good quality reads at 70bp and longer; SNPs have to be present on the same sequencing read as a CpG for ASM analysis. SNP coordinates were treated identically to the CpG coordinates. If allelic frequency information was available for a SNP site, I included it in the coordinate list. I then associated all coordinates with RefSeq genes. I also noted where each CpG coordinate was located within a gene. For example, if a CpG coordinate was located within an exon, I marked it as such. In order for a CpG coordinate to be marked in a promoter or 3' region, it had to be less than 2 kb away from the gene's TSS or stop site. Regions containing labeled CpGs and SNPs that were separated by at least 500bp were broken up into separate regions. Regions that contained T/C SNPs were marked so that padlock probes would target the appropriate strand. This is important because after BSC, a T/C SNP site becomes undistinguishable

unless the cytosine is always methylated. In these cases, the opposite strand, which contains an A/G SNP site, is targeted so that one can distinguish the alleles.

Using a probe selection algorithm specifically designed for padlock probes, I then produced a set of candidate padlock probes that covered all targetable CpGs and associated SNPs. The probe selection optimizes target region coverage and the quality of padlock probes. Multiple strand-alternating padlock probes are produced to target regions that cannot be targeted by a single padlock probe; padlock probes whose target regions overlap cannot be on the same strand since they would interfere with each other. Since padlock hybridization occurs on ssDNA, padlock probes targeting opposite strands would not interfere with each other, even if their target regions overlapped. The T_m of both targeting arms of each candidate probe are calculated and only padlock probes where both T_m values fall within a specified range are accepted. This T_m characteristic helps predict the hybridization capability of the padlock probe and ensures that the padlock probes bind more uniformly to their target regions.

The folding energy of each candidate padlock probe sequence is also predicted via a program called Unafold(*205*), which predicts the folding energy of a nucleotide. Computationally derived folding energy values were found to correlate with the padlock probes' target region amplification efficiencies. Folding energy calculations help take into account the formation of a secondary structure within a nucleotide sequence. Formations of secondary

structures within a padlock probe set would introduce variability since some padlock probes would form secondary structure structures that hindered hybridization efficiency at targeted regions. The presence of a CpG island in a targeting arm leads to the generation of additional padlock probes so that a padlock probe's targeting arms will form a perfect match regardless of the CpG methylation state. Since a CpG in a targeting arm's hybridization region leads to additional padlock probe generation without any benefit in coverage, such regions that contain a minimum number of CpGs are preferred. SNPs within targeting arms also present hybridization challenges and targeting arm regions that overlap with candidate SNP sites are also avoided.

Due to the large number of regions for which candidate padlock probes were needed and the computational complexity involved in the selection of padlock probes (e.g. T_m calculations, folding energy calculations, creation of overlapping padlock probes for large regions, etc.), this process had to be adapted so that it could run in parallel. Since each targeted region is independent from every other padlock probe region, this padlock probe selection process can be run separately on each region. I utilized this fact to create a padlock probe pipeline that could run on 90 nodes simultaneously and thusly accelerate this process by 90x. A probe selection process that would take months on a single CPU was reduced to a few days. This was especially helpful since many padlock selection runs were required as I adjusted parameters (e.g. minimum T_m, optimal targeting arm length, etc.) and modified the list of targeted regions.

Once the padlock probes were generated, I associated the targeted regions with CpGs of interest and ensured that the coverage was acceptable (85% or higher). I also used SOAP to check the uniqueness of the targeting arms' sequences(*206*). I required each padlock probe to hybridize to a single location on the genome. Otherwise, a padlock probe could target multiple areas and introduce noise into the results. Candidate padlock probes were then ranked based on their targeted CpG count, targeted SNP count, and regions associated with RefSeq genes. The CpG count is important to LD and DNA methylation pattern analyses, the presence of SNPs is important for ASM analysis, and the RefSeq gene is important for identifying relationships between DNA methylation and protein coding gene expression.

I ranked data sets for which padlock probe coverage was limited (i.e. less than 85%). Regions identified by Stamatoyannopoulos' and Feinberg's groups were considered to be of greatest interest. I thusly included all padlock probes covering these regions into the Cpg220k data set; ranking was unnecessary. However, the Salk DMRs, for which there are millions of CpGs, were ranked since I included only a small percentage of padlock probes covering those regions into the Cpg220k probe set. Finally, the top candidates from the ranked padlock probe lists (i.e. Salk DMRs, AML DMRs, predicted enhancers, and imprinting control regions) were chosen and added to the Cpg220k probe set.

2.7 Cpg220k Statistics

The Cpg220k probe set contains 220,000 padlock probes, uniquely covers 19,205,634 bp, and targets 119,474 unique regions. Cpg220k covers 620,708 CpGs and 105,206 CpGs are covered on both strands. Double-stranded coverage can be used to verify the experiment by comparing stranded methylation frequency values and double-stranded coverage also improves SNP calling. This probe set also includes 127,484 SNPs. 85% of CpGs within Feinberg's DMRs are covered (117,376 CpGs covered), 89% of CpGs within Stamatoyannopoulos' data set are covered (251,312 CpGs covered), 2% of CpGs within Ecker's DMRs (112,585 CpGs covered), 19% from the predicted enhancers data set (30,662 CpGs covered), 18% from the acute myeloid leukemia data set (16,161 CpGs covered), and 27% of CpGs located proximal to or within imprinted genes (9,353 CpGs covered). Table 2.7.1, Graphs 2.7.1-2.7.3, and Figure 2.7.1 describe additional details of the Cpg220k probe set.

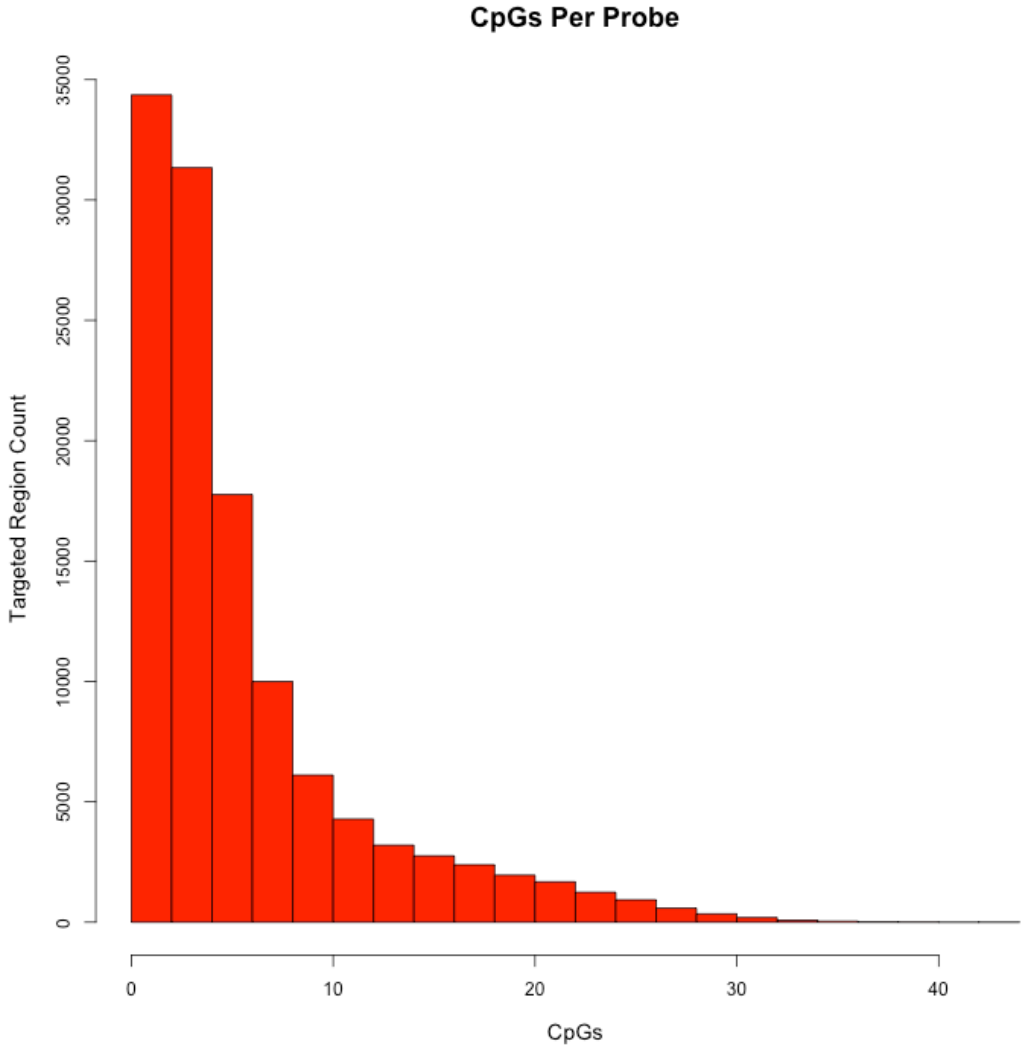
Besides expanded coverage, additional design changes were made to improve the Cpg220k data set. I extended the optimal targeting arm lengths to 27bp so that each arm would likely hybridize at a single location and thusly reduce any off-target sequencing artifacts. In previous probe sets the target length was variable. For example, target regions could be between 175 and 225 bp in the Cpg30k probe set. The variability in target region sizes lead to larger smears in PAGE gels during size selection steps. Smears can make it more difficult to remove all of the target sequences and thusly prevent the amplification of desired regions. Large smears can also lead to the inclusion of

sequences that are not targeted regions. To minimize the target regions' smear area during electrophoresis, all target regions in the Cpg220k probe set are 180 bp long. Since all target regions are of the same length, they will appear as a single tight band in PAGE gels, making their excision and identification from other sequences easier.

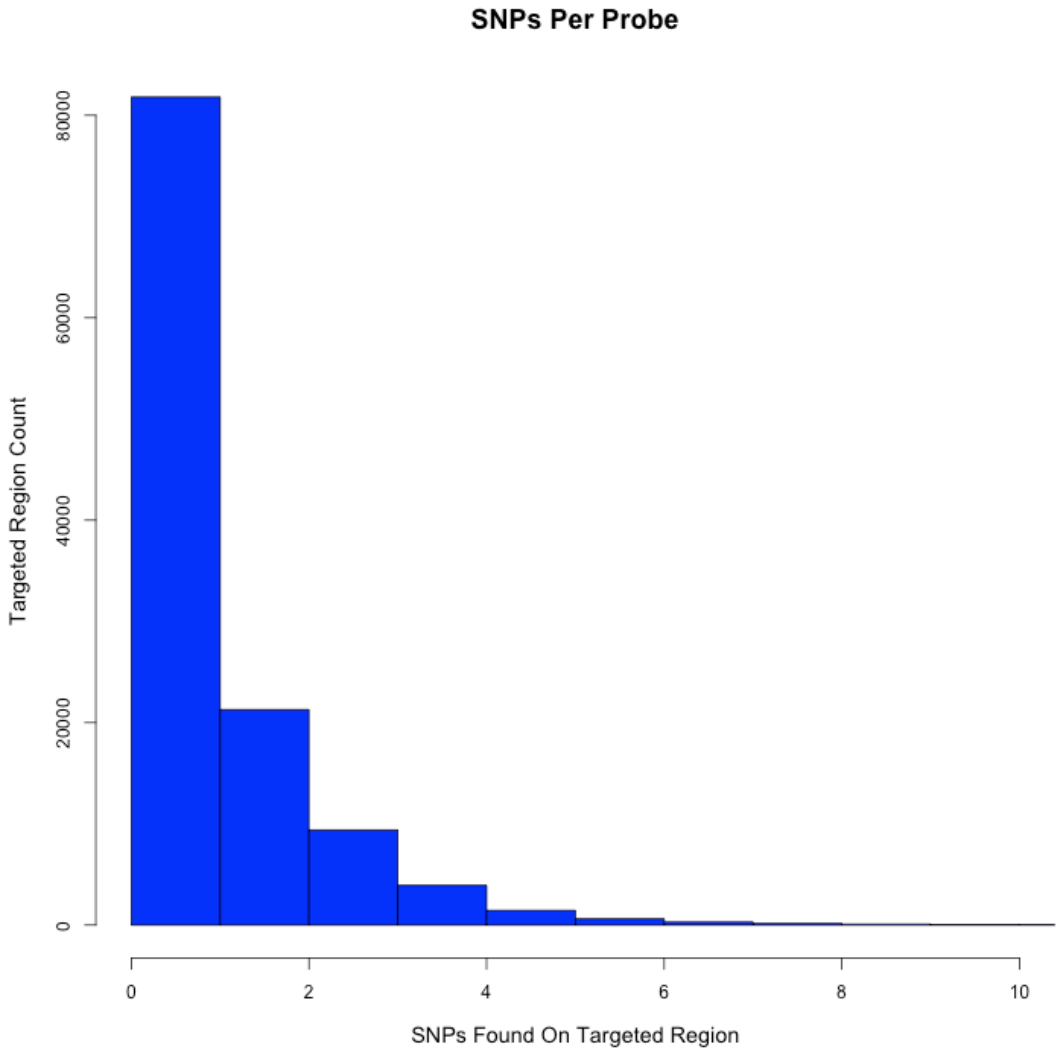
Table 2.7.1 CpG coverage by probe set. This table shows the number of CpGs covered by bisulfite sequencing padlock probe sets. Cpg30k is the oldest and was used to demonstrate the BSPP method(*38*). The Cpg30k probe set focused on CpG islands in chr12 and chr20 while the newer Cpg97k set focused on DMRs published by Feinberg's group(*40, 41, 186*). The number in each of probe set name represents the number of probes in that set (e.g. Cpg30k contains about 30,000 padlock probes).

Probe Set Name	Number of CpGs Covered
Cpg30k	129,175
Cpg97k	245,221
Cpg220k	620,491

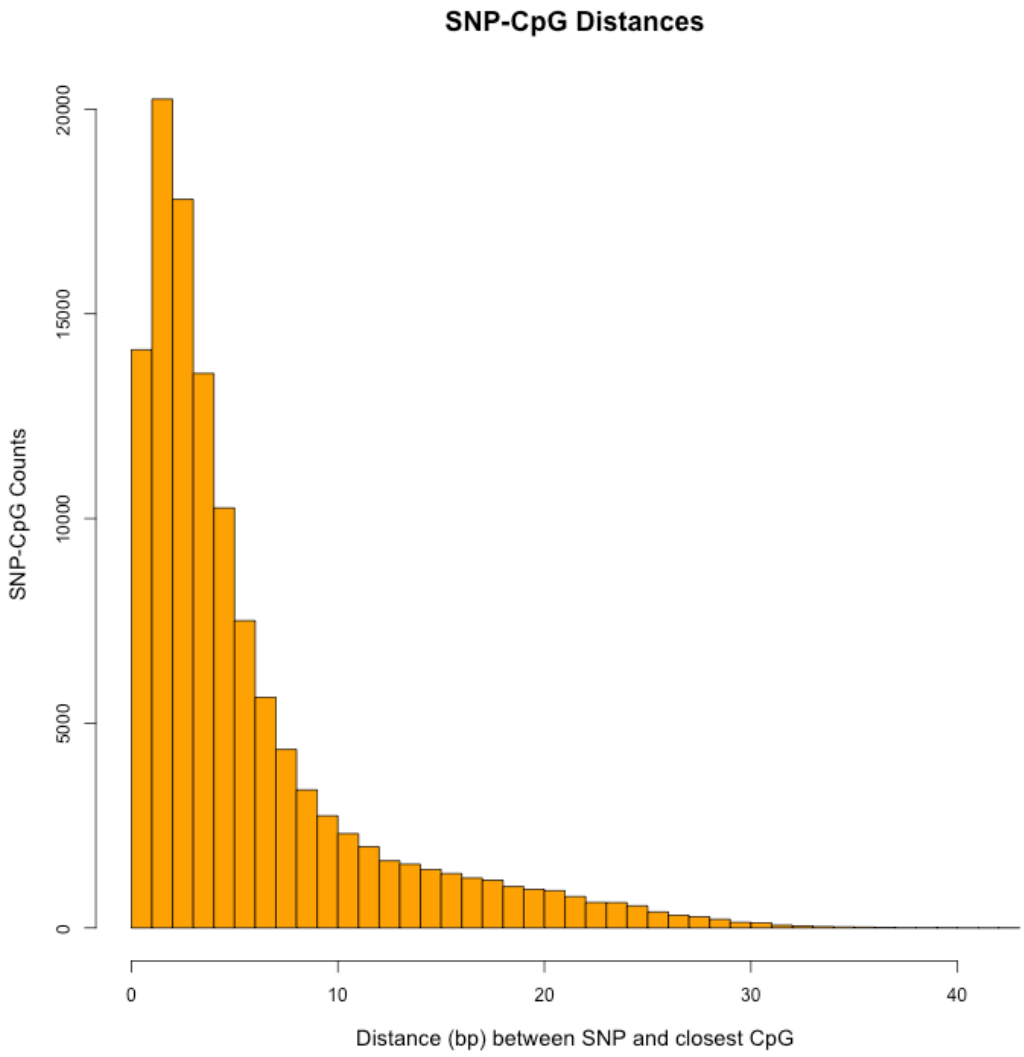
Graph 2.7.1 CpGs per probe count. This histogram shows the number of CpGs covered per probe. The y-axis represents the density of targeted regions associated with the CpG range indicated on the x-axis. Since most the targeted regions are not CpG islands, the histograms reflects that between 25% and 30% of targeted regions contain less than 4 CpGs. Each targeted region spans 180 bp.



Graph 2.7.2 Padlock probe SNP density. This histogram shows the number of SNPs contained on a padlock probe. Most targeted regions do not cover a candidate SNP region. In order for a SNP to be interesting in methylation analysis, a sequencing read must cover the SNP site and at least one CpG site.



Graph 2.7.3 SNP-neighboring CpG distances. This histogram shows the distance between a candidate SNP site and its closest CpG (bp). 70-100bp reads are becoming common in short read sequencing, which means that SNP-CpGs distances less than 70-100bp will be found on the same read. This information is important for ASM analysis since it requires a SNP and at least one CpG to be present on the same read. The Cpg220k probe set is the first BSPP (bisulfite sequencing padlock probe set) that is designed to capture SNPs and CpGs together.



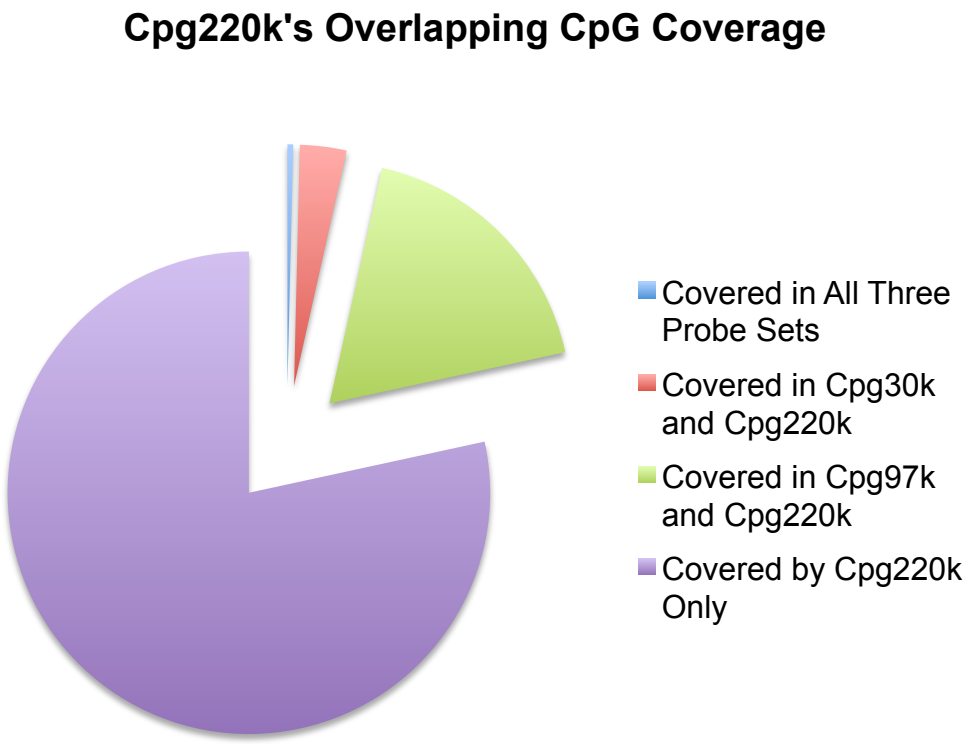


Figure 2.7.1 Cpg220k's overlapping CpG coverage. This pie chart shows the dramatic increase in coverage of the Cpg220k probe set. Since most of the CpGs covered in the Cpg30k probe set showed little methylation change across the investigated cell types (fibroblasts, lymphoblasts, and pluripotent cells), most of these CpGs were not targeted in the newer data sets. The Cpg220k and Cpg97k overlap much more significantly because both probe sets are larger than Cpg30k and both sets include DMRs found by the Feinberg group

Chapter 3 Intercellular Methylation Analyses Using Bisulfite Padlock Probes

3.1 Background and Experimental Validation

The bisulfite sequencing padlock probe (BSPP) technique performed in Kun Zhang's lab was first published using the Cpg30k probe set. The goal of this study was to validate the BSPP capture method by interrogating CpGs across multiple fibroblast, ES (embryonic stem cell), and iPS (induced pluripotent stem cell) cell lines. My work on this project involved finding methylation patterns in these cell lines and discovering the biological meaning for such patterns.

The Cpg30k probe set targeted 2,020 CpG islands across chr12 and chr20, 237 promoters in eight ENCODE regions, and the 4 kb region centered around the TSS of 26 genes associated with development or pluripotency. Validation of the BSPP capture method involved the reproducibility of methylation frequencies across IMR90 replicates, the consistency of stranded methylation frequency values (i.e. whether the methylation frequency of a CpG on the Watson strand matched the methylation frequency of the same CpG on the Crick strand), and the comparison of the Illumina sequencing methylation frequency results with an independent sequencing method (i.e. Sanger sequencing). The IMR90 replicates exhibited a strong correlation, R=.97, and the strand specific methylation comparison exhibited a strong correlation of

R=.987. The strong correlation of the IMR90 replicates demonstrated that the entire experiment could be reproduced with little variability, either biological or experimental, and the strand specific validation showed that there is only a small error involved within the experiment itself. The strand specific CpG comparison acts as an intra-experiment variability marker. After bisulfite conversion (BSC) the DNA strands are no longer reverse complementary and thusly act as independent sequences. The later discovery that most of the targeted CpGs are unmethylated in the Cpg30k dataset, which means most CpGs as well as non-CpG cytosines are converted to uracils thusly destroying base pairing, further supports the independence of the BSC Watson and Crick strands. Finally, the Sanger sequencing results matched the methylation frequencies values from Illumina sequencing with a correlation of R=.975. These validation steps demonstrate that the Cpg30k probe set reliably captured methylation states from targeted CpGs.

3.2 Cell Lines Used in the Experiment

The methylation analysis focused on finding methylation signatures related to pluripotency and development in 11 cell lines. 4 of the investigated cell lines were induced pluripotent stem cell lines. Another 3 were the differentiated fibroblast cell lines from which the iPS cells were derived. Three embryonic stem cell lines were used and also a single Hybrid1 cell line, which consisted of a HUES6 nucleus fused with a BJ nucleus. Given the range of differentiated cells, embryonic stem cells, and experimentally manipulated

pluripotent stem cells, significant methylation-related features found in these datasets could be linked to phenotype (e.g. pluripotent or differentiated) or to a specific cell line.

Four iPS cell lines and their fibroblast cell line sources were included in this study. The three fibroblast cell lines are BJ (neonatal human foreskin fibroblasts), IMR90 (fetal lung fibroblasts), and hFib2 (adult male fibroblasts). From these three fibroblast cell lines, four iPS cell lines were produced, BJ-iPS11, BJ-iPS12, hFib-iPS4, and IMR90-iPS. Multiple methods for iPS transformation exist. Since there is strong social pressure for researchers to derive pluripotent stem cells without the destruction of embryos, there is great interest in the development of an iPS protocol that optimizes efficiency and viability (e.g. minimize cancer risks). The four studied iPS cell lines were derived via three different transformation protocols: (1) IMR90-iPS was reprogrammed with Oct4, Sox2, Nanog, and Lin28 transcription factors(*106*), (2) hFib2-iPS4 was reprogrammed with Oct4, Sox2, Klf4, and Myc(*207*), and (3) BJ-iPS11 and BJ-iPS12 were transformed with Oct4, Sox2, Nanog, Klf4, and Myc via an inducible promoter(*208*). The inducible promoter technique allows those five transcription factors to be expressed upon addition of a chemical. This allows researchers to differentiate iPS cells into various cell types and then reprogram them back to iPS cells by adding a single chemical.

Representing another pluripotency reprogramming technique, a pluripotent Hybrid1 cell line that contains fused BJ and HUES6 nuclei was also investigated. The Hybrid1 cell line is a tetraploid since it contains 2

chromosomes originally from a BJ nucleus and 2 chromosomes originally from a HUES63 nucleus. Methylation frequencies computed for this cell line will thusly be an average over four chromosomes. The Hybrid1 cell line is of research interest since it represents yet another technique for reprogramming a fibroblast cell to a pluripotent state. However, this cell line offers little medical potential since it requires a HUES nucleus for pluripotency transformation and the biological consequences of a tetraploid human cell are not well understood.

Serving as a pluripotency standard, three human embryonic cell lines were examined in the study. HUES12 was first published in 2004 and it was derived through the immunosurgical removal of the inner cell mass from a human blastocyst(*209*). HUES43 and HUES62 cell lines were derived using an updated protocol that replaced immunosurgery with a laser, and these cell lines were published in 2009(*210*). Although the three stem cells show similarity regarding presence of pluripotency markers, researchers have found different human embryonic stem cells have different differentiation biases(*211*). Comparing these gold standard cell lines to the iPS and Hybrid1 cell lines, one can observe methylation patterns that offer clues into any characteristic differences in the transcriptional regulatory program amongst the cell types.

3.3 Overview of Methylation Across Cell Lines

Looking broadly at chromosomes 12 and 20, where the vast majority of Cpg30k targeted regions lie, the methylation signatures appear to be generally consistent across the 11 cell types (Figure 3.3.1). Between 58,000 and 73,000 CpGs per cell line were covered by at least 10 reads in the Cpg30k BSPP experiment and these CpGs were used in the methylation signature analysis (Table 3.3.1). From Figure 3.3.1, methylation appears to be in blocks. Unmethylated regions flank regions of strong methylation. However, the Cpg30k BSPP is limited since it is misses CpG methylation information outside of CpG islands on chr12 and chr20.

Many of the targeted CpG islands appear unmethylated consistently across all cell types (Figure 3.3.2). A histogram showing the distribution of methylation frequencies further illustrates this consistent demethylation of CpG islands (Graph 3.3.1). Mentioned in Chapter 2, the presence of targeted regions whose DNA methylation signatures remain static is uninteresting when looking for ways to discriminate between cell types. Ultimately, I am discovering differences in methylation patterns and directly or indirectly linking them to differences in biology (e.g. phenotypic differences such as pluripotency and differentiated cell types). Although many regions do not show broad cell type specific methylation patterns, there are some regions that do show unique methylation patterns. Figure 3.3.3 shows a ~15 Mb region in chromosome 20 that shows short but distinct fibroblast- and pluripotent-specific methylation signatures. This small region of differential methylation

overlaps with several genes (e.g. C20orf26, PLK1S1, CRNKL1) and the presence of the DNA methylation likely has significant biological outcomes for many of these overlapped genes. Given the presence of cell type specific methylation regions, the next step is to examine the discriminatory power of these regions and find biological reasons for these methylation differences.

Table 3.3.1 Read mapping statistics for the Cpg30k BSPP Experiment.

Cell line	Raw reads	Mapped Reads	CpG sites covered ≥ 10 reads	Mean Coverage	Non-CpG Cytosines (%)
BJ	10,853,272	2,057,753	67,470	71	0.01
BJ-iPS11	12,276,017	2,240,958	69,454	77	0.012
BJ-iPS12	10,539,126	2,179,019	70,741	74	0.011
hFib2	14,408,007	2,540,405	60,007	96	0.011
hFib2-iPS	14,805,000	2,173,844	58,531	80	0.014
IMR90	14,795,564	2,896,881	59,950	113	0.011
IMR90.2	9,892,238	2,051,461	55,966	72	0.011
IMR90-iPS	10,985,469	2,473,997	66,424	58	0.013
Hybrid1	9,693,455	2,913,639	68,849	65	0.014
HUES12	11,360,235	3,760,194	72,463	118	0.015
HUES42	11,185,368	3,671,564	67,926	125	0.012
HUES63	14,315,301	2,472,893	73,196	97	0.013

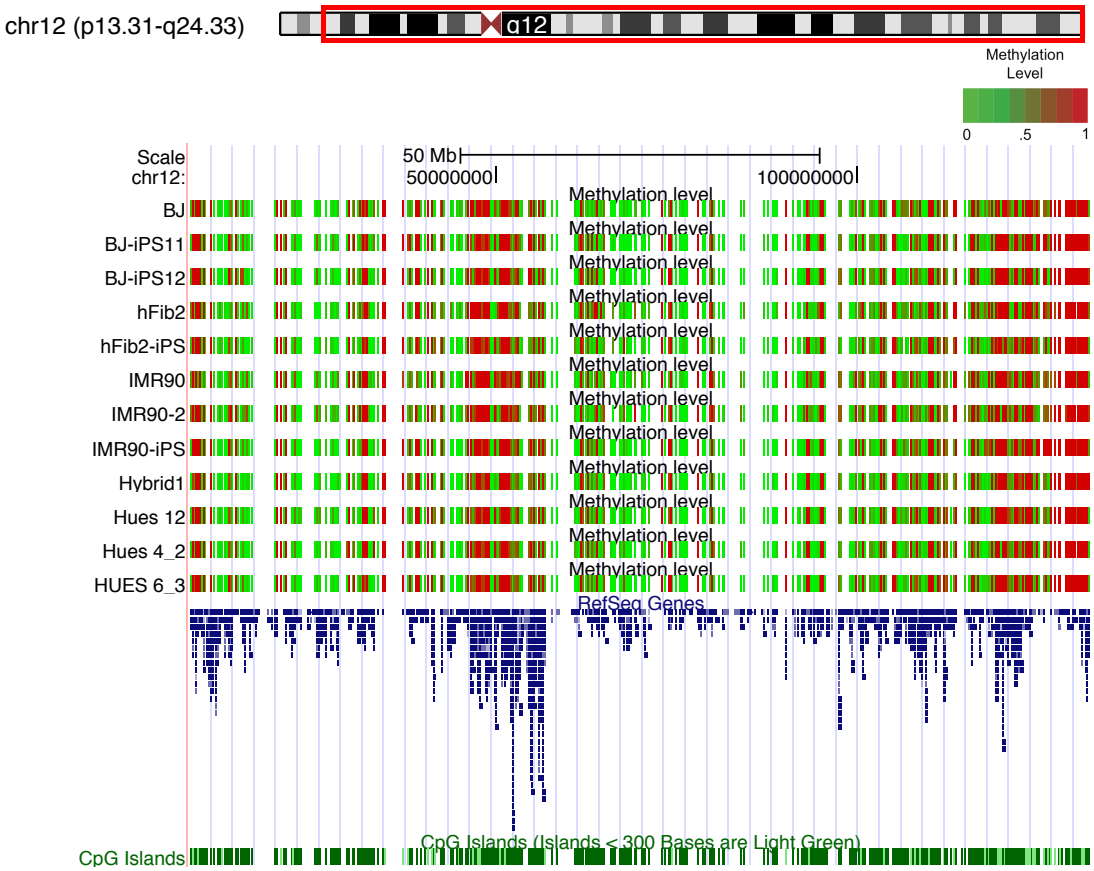


Figure 3.3.1 A ~100 Mb window from chr12. In the chromosome ideogram (top), the red rectangle specifies how much and which region of the chromosome is displayed below. Green bars indicate unmethylated CpGs and red bars indicate methylated CpGs. A this scale, the methylation landscape appears to be similar across cell lines.

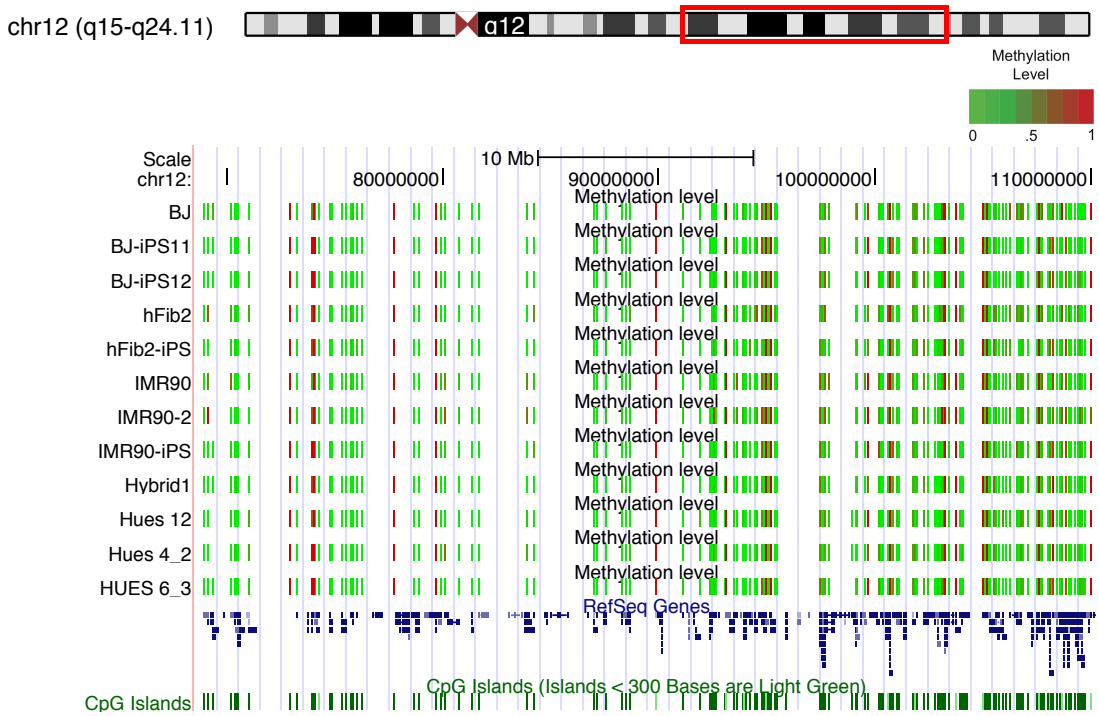
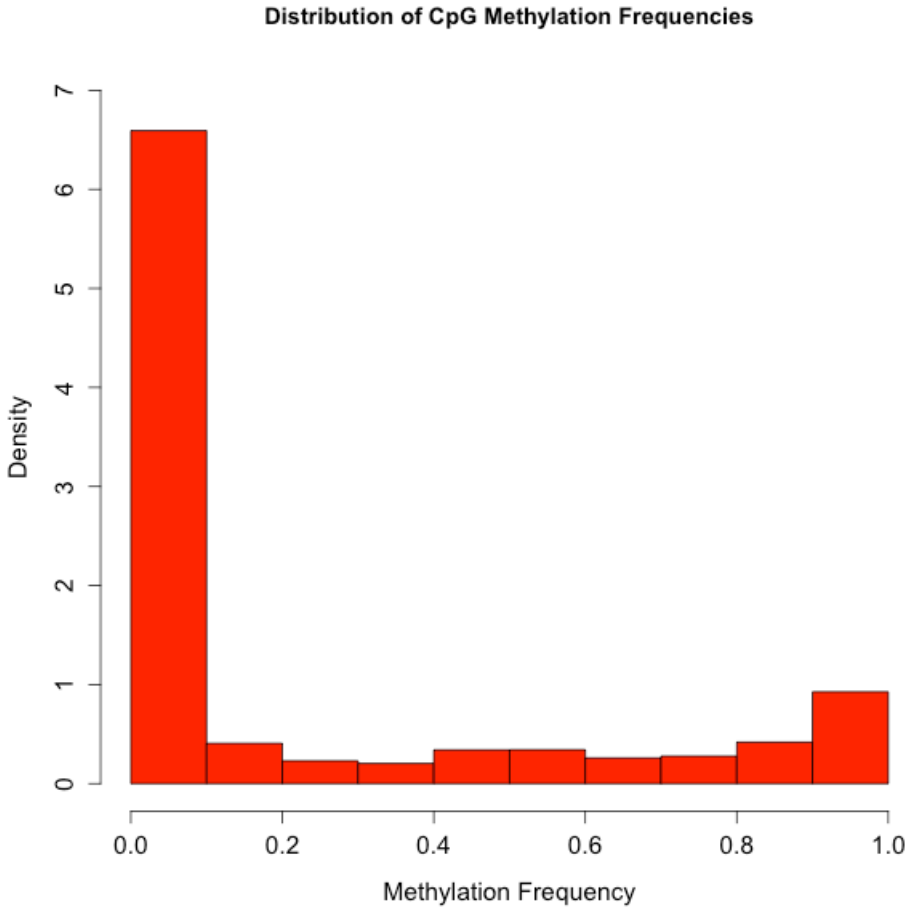


Figure 3.3.2 A smaller section of chr12 (~35 Mb) reveals that many CpG islands are unmethylated.

Graph 3.3.1 This graph shows the distribution of CpG methylation frequencies across all cell lines. The area underneath the bars sums to unity. The distribution shows that over 60% of the CpGs targeted by the Cpg30k BSPP are unmethylated while only about 10% are completely methylated.



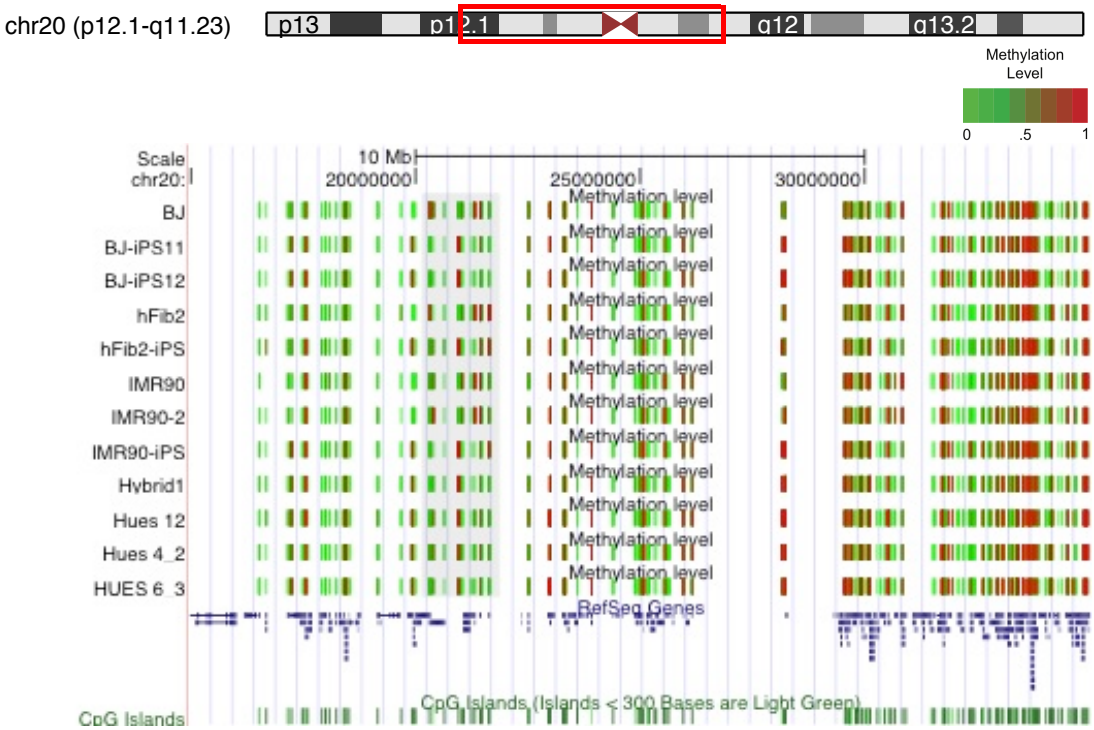


Figure 3.3.3 This chromosomal region reveals cell type specific methylation patterns on a large scale (~15 Mb window). One such region is within the darkened rectangle between 20 and 25 Mb. Fibroblasts share a methylation signature in this region that visibly differs from pluripotent cell types. The iPS cell types have signatures that closely match the HUES cell lines in this region.

3.4 Cell Type Specificity of DNA Methylation

Although most of the targeted CpG sites were consistently unmethylated across all cell types, cell types clustered together based on their methylated frequency profiles (Figure 3.4.1). Unsurprisingly, the IMR90 replicates clustered closely together while the BJ and hFib2 cell lines showed more methylation variability. This variability is not surprising since IMR90 cells originated from fetal lung tissue, BJ cells originated from neonatal foreskin tissue, and hFib2 came from adult dermal tissue. Despite this variability, the fibroblast lines have methylation signatures that are noticeably different than the pluripotent cell lines.

The embryo-derived pluripotent cells, HUES63, HUES42, and HUES12, clustered very closely together showing similar variability to the IMR90 replicates. This demonstrates a methylation state homogeneity of targeted CpGs across the HUES cell lines, which supports their use as a standard of pluripotency.

The iPS cell lines clustered much more closely to the HUES cell lines than fibroblasts, supporting their identity as pluripotent stem cells. But the iPS cell lines clustered separately from the HUES cell line. The difference in similarity between the fibroblasts and the iPS transformed fibroblasts demonstrates that the iPS cell biology, as revealed by DNA methylation, closely matches the embryo-derived pluripotent cell lines. In addition, the methylation signature similarity of all iPS cell lines to HUES cell lines

demonstrates that the three transformation methods do not yield vastly different methylation profiles. However, the two BJ-iPS cell lines do cluster most closely together, which may reveal a fibroblast memory effect in the iPS cell line or that differences in transformation methods yield small differences in DNA methylation signatures. Unlike the iPS cell lines, the Hybrid1 cell line is indistinguishable from the HUES cell lines. Since the Hybrid1 cell contains fused BJ-HUES6 nuclei, the closer similarity of Hybrid1 to HUES cell lines relative to iPS cells hints at methylation dependencies on transformation methods.

Comparison of the methylation frequency distributions between fibroblasts and pluripotent cells reveals higher levels of methylation in the pluripotent cell lines. This is observed by looking at a Quantile-Quantile plot, which involves: (1) taking CpG sites covered by all cell lines, (2) averaging the methylation frequencies per CpG across cell lines within the same cell type group (i.e. fibroblasts, ES cells, iPS cells), (3) ranking the methylation frequencies for each group separately, and (4) plotting the methylation frequencies of two groups based on rank (Graph 3.4.1). If the distribution of methylation frequencies were exactly the same for both cell lines, the points would match the y=x line. However, the Q-Q plot shows that in intermediate methylation frequency values (i.e. values between .2 and .8), pluripotent stem cells appear to have higher methylation frequencies at the same rank than fibroblasts. For example, as one descends down the ranked list of fibroblast methylation values, he/she finds that a .4 methylation value corresponds to a

.6 methylation frequency value at the same rank in pluripotent cells. More interestingly, pluripotent cells are slightly more methylated than ES cells based on the Q-Q plot. Overall, pluripotent cells tend to be more methylated than fibroblasts.

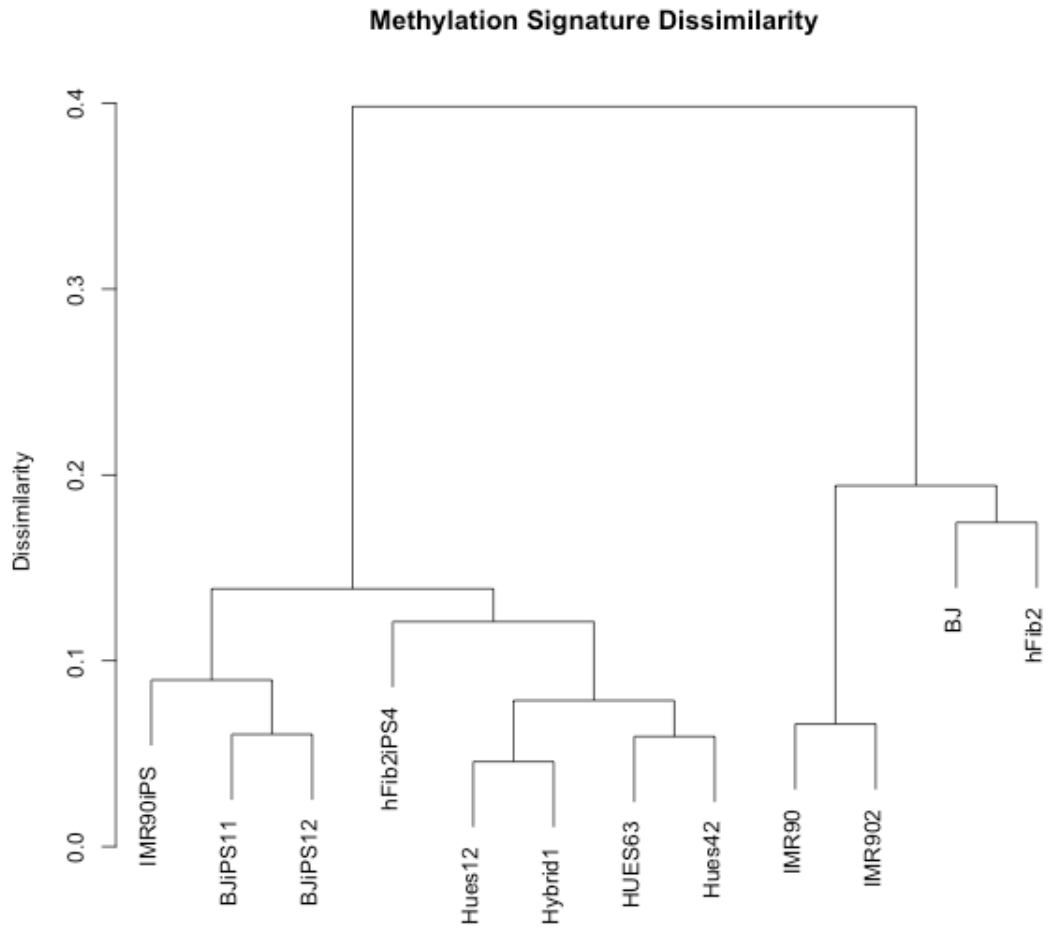
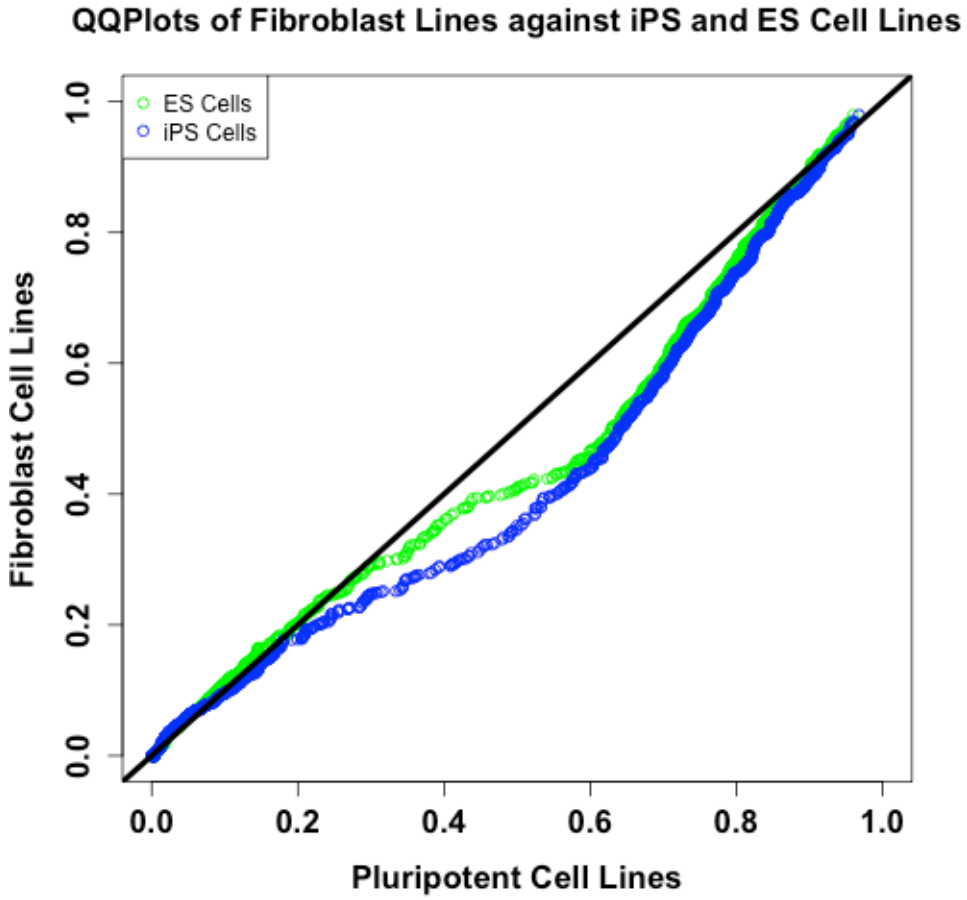


Figure 3.4.1 This graph depicts the dissimilarity between cell lines based on the Cpg30k targeted CpG methylation singatures. Dissimilarity is the euclidean distance between correlations. The fibroblast lines, IMR90, BJ, and, hFib2, cluster closely together while the iPS, HUES, and Hybrid1 cell lines cluster closely together. Within the pluripoent cell type group, the iPS cells do not cluster as tightly together as the HUES cell lines.

Graph 3.4.1 This plot compares the methylation states of CpGs that were covered in all 12 samples. The methylation values were averaged per cell type group (i.e. ES cells, iPS cells, and Fibroblast cells). These averages were ranked and plotted rank wise. For example, the highest methylation frequency value for the fibroblast group (y-axis) was plotted with the highest methylation frequency value for the pluripotent cell group (x-axis). The graph shows that for CpGs with intermediate methylation values tended to be more methylated in pluripotent cell lines. Additionally, these regions tended to be more methylated in iPS cells than in HUES cell lines.



3.5 DNA Methylation in TSS Regions

Given that the overall methylation profiles distinguish between cell types, more detailed examinations of targeted CpG regions illuminates the biological meaning behind these differences. When looking at the 26 genes whose TSS regions were covered by the Cpg30k probe set, certain genes associated with development and pluripotency show cell type specific methylation patterns (Figure 3.5.1). The *POU5F1* TSS region, which encodes the Oct4 protein (i.e. master regulatory of pluripotency), is more methylated in the fibroblast cell lines than in the pluripotent cell lines. However, *NANOG* and *SOX2*, which encode transcription factors associated with pluripotency, do not show methylation patterns that clearly discriminate between the pluripotent and differentiated cell types. The TSS region of *DNMT3B*, which is associated with de novo methylation in pluripotent cell lines, does show significant methylation in the fibroblast cell lines relative to the pluripotent ones.

Though *POU5F1* and *DNMT3B* show cell type specific methylation patterns, they are not as simple as completely unmethylated or methylated (Figures 3.5.2-3). In the pluripotent lines, *DNMT3B* is mostly unmethylated, with the vast majority of methylation frequencies less than .15 but in the fibroblast, the methylation frequencies assume a broader distribution. Most of the CpGs in this TSS region are either methylated at around .7 or .05 across the fibroblast cell lines. Few CpGs are strongly methylated (i.e. methylation frequency > .9). Across the fibroblast cell populations used in the BSPP

experiment, the majority of cells contain methylated CpGs in the *DNMT3B* region but there exists a small subpopulation of cells, or more unlikely alleles, where these CpGs are also unmethylated. *POU5F1* shows a similar pattern where its TSS region is mostly unmethylated (vast majority of methylation frequencies < .2) in pluripotent cells but the methylation frequencies in the fibroblast cells are more broadly distributed in this region. Gene expression correlates negatively with methylation in the *DNMT3B* and *POU5F1* TSS regions. Gene expression is reduced over 10-fold in fibroblasts, which also show higher levels for CpG methylation in both TSS regions (Table 3.5.1). These two TSS regions demonstrate that shifts in methylation patterns can correlate with dramatic changes in expression level.

Similar to *POU5F1 and DNMT3B, NANOG* and *LEFTY2* also undergo significant expression change upon differentiation but their respective methylation patterns do not change as significantly (Figures 3.5.5-6). *LEFTY2* undergoes an expression change similar to *POU5F1* upon differentiation but the methylation pattern in its TSS region is similar across the three cell types. *NANOG* is also only highly expressed in pluripotent cell lines. The methylation states around its TSS show slightly more methylation in the fibroblast cell lines, but it is not as clear as with *POU5F1* and *DNMT3B*. It is likely one or both of the following factors: (1) regulatory mechanisms other than DNA methylation contribute significantly to these genes' transcription in this region and (2) regulatory regions outside the TSS regions are contributing to these genes' transcription. In the case of (2), targeting those distal regulatory regions

may reveal DNA methylation changes similar to the TSS regions of *POU5F1* and *DNMT3B. FGF7* is a gene that is significantly more expressed in fibroblasts. Its methylation profiles show that its TSS region is more unmethylated in fibroblasts than in pluripotent cells (Figure 3.5.4).

Two known imprinted genes, *SNRPN* and *MEG3*, show little methylation signature changes upon differentiation. Imprinted genes are allele specifically expressed genes and the allelic preference if parental specific. Chapter 1 describes mechanisms that underlie imprinting and the resulting phenotypes from improper establishment of imprinting regions. Imprinted regions are regulated by imprint control elements; if imprint control elements are removed, the imprinting pattern is lost. The TSS regions around *SNRPN* and *MEG3* show that the majority of CpGs are fuzzily methylated, which means the CpGs contain a methylation frequency between .25 and .75. In the case of imprinting genes, the fuzzy methylation values are caused by averaging over an unmethylated and a methylated allele. Since the methylation frequencies do not significantly change across cell types, the imprinting patterns appear to be already established in pluripotent cells. Reprogramming also does not affect these patterns.

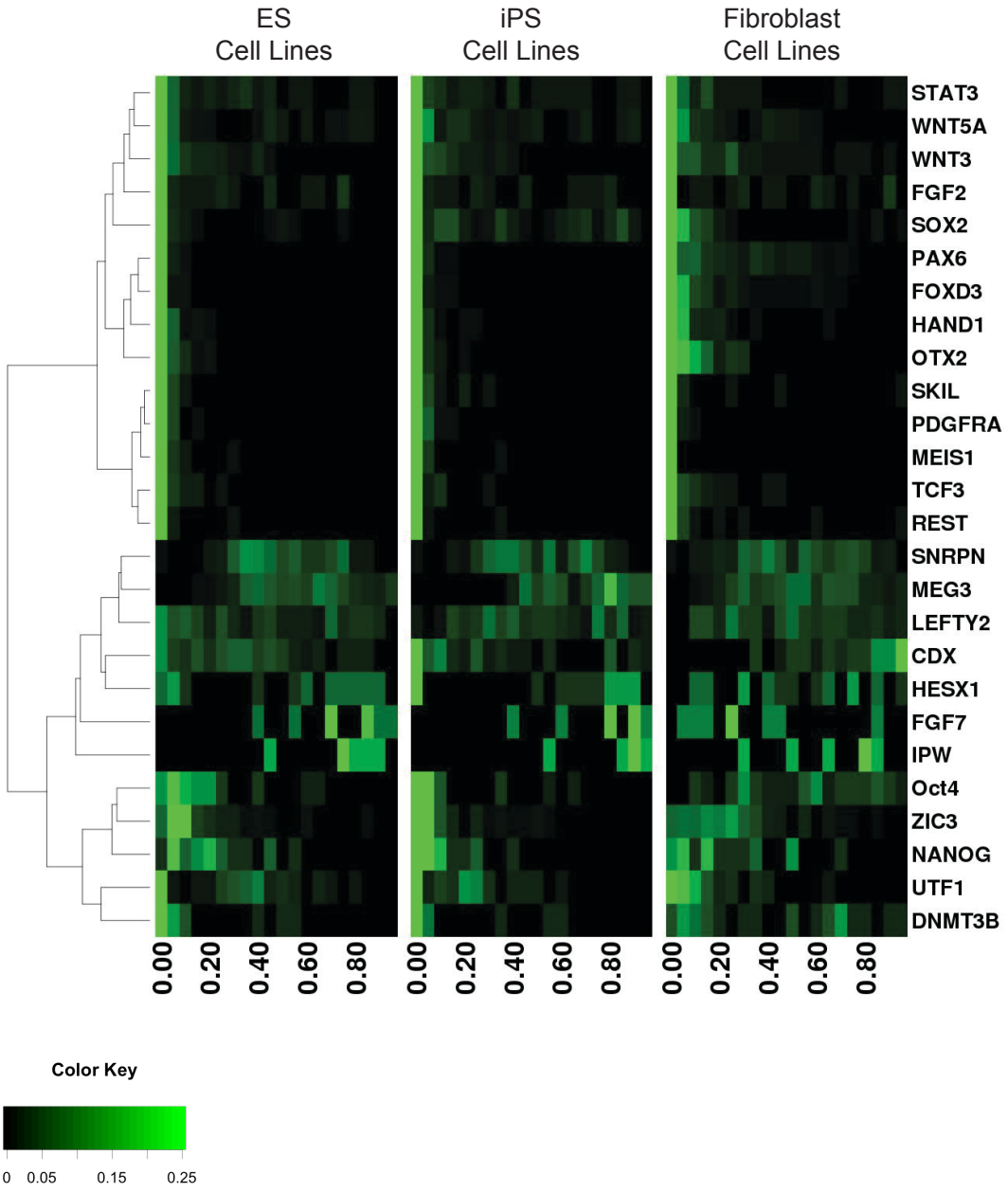


Figure 3.5.1 This figures shows the methylation profiles across 26 genes ([+2kb,-2kb] from TSS). CpGs are placed into 20 bins based on their methylation frequency values. For example, if a methylation frequency is between .20 and .25, it is placed in bin 4, which is labeled as 0.20. The color of each bin represents the fraction of probes in the TSS region that fall within the bin. The higher the fraction, the brighter the green color. For example, if all probes within a TSS were contained in the .60 bin, there would be a single bright green mark at the .060 bin; all other bins would be black. The color saturates at .25 so that subtle methylation changes are visually distinguishable.

Table 3.5.1 This table shows the log2 expression values for referenced genes. Note that HUES6 was not a cell sample used in the methylation analysis but it is considered as a reference ES cell. All expression values were quantile normalized.

RefSeq Gene	Gene Name	HUES6	Hybrid1	BJ	BJ-iPS12	IMR90-iPS	hFib2
NM_006892	***DNMT3B***	13.4	13.3	6.0	8.5	10.2	6.2
NM_002701	***POU5F1***	13.4	13.4	5.6	10.5	10.8	5.6
NM_024865	***NANOG***	12.8	12.6	5.3	9.7	9.5	5.6
NM_003240	***LEFTY2***	13.6	12.9	5.8	10.5	9.2	5.9
NM_002009	***FGF7***	5.1	5.2	6.7	5.4	4.9	9.0

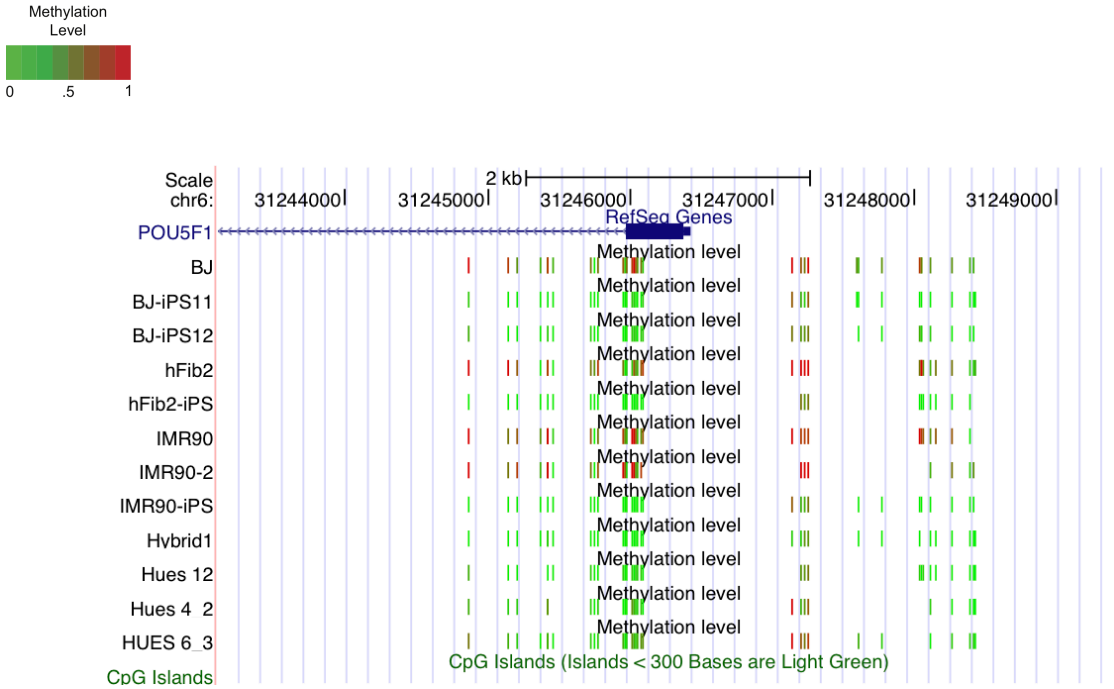


Figure 3.5.2. This figure shows the CpG metylation levels around the *POU5F1* TSS region. *POU5F1* is expressed in pluripotent cells and is less methylated around the TSS.

Figure 3.5.3. This figure shows the CpG methylation levels around the *DNMT3B* TSS region. *DNMT3B* is expressed in pluripotent cells and, like *POU5F1,* is less methylated around the TSS. Unlike in the *POU5F1* TSS region where the cell type specific methylation pattern is uniformly spread throughout, the unique mehylation pattern in *DNMT3B* weakly appears in the promoter region and more obviously about 2 kb downstream of the TSS. The covered CpG island does not show cell type specific methylation.

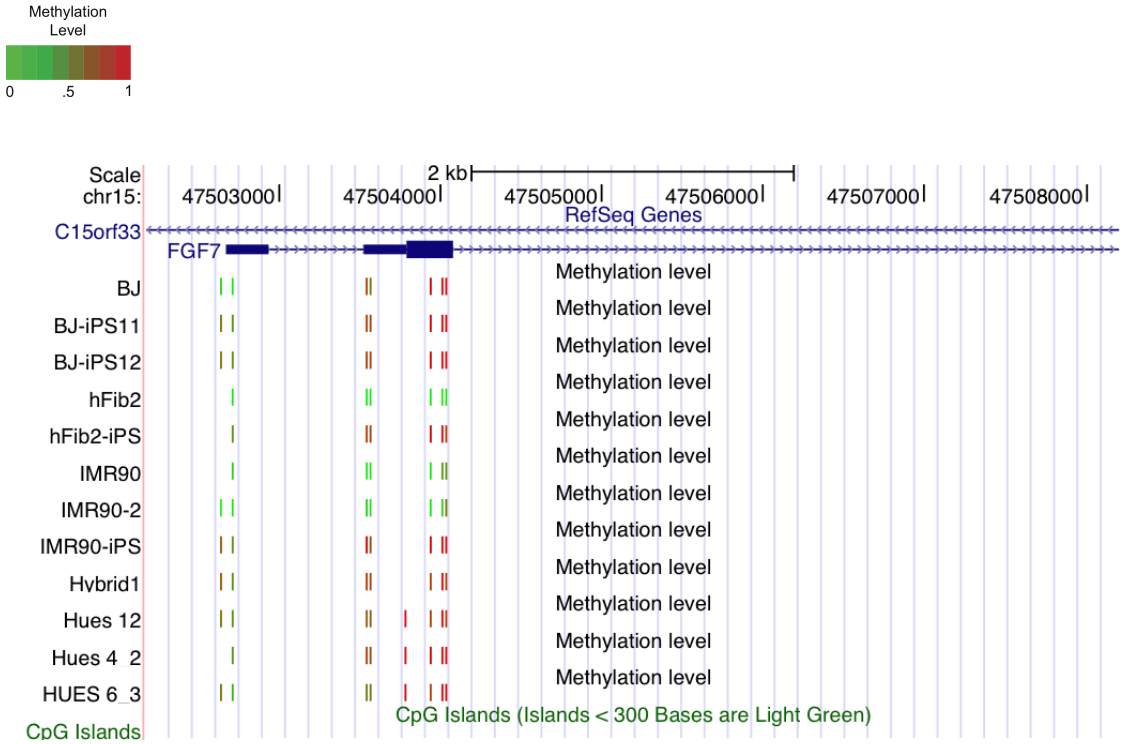


Figure 3.5.4 Similar to *DNMT3B* and *POU5F1*, *FGF7* is more expressed in cell lines that exhibit less methylation around the TSS. The pluripotent cell lines show the most methylation around the TSS, while BJ shows moderate methylation, and the IMR90 and hFib2 show the least methylation. The expression follows this progression with hFib2 showing the higest level of *FGF7* expression, BJ showing moderate *FGF7* expression and the pluripotent cell lines showing the lowest levels of *FGF7* expression.

Figure 3.5.5 Contrary to *DNMT3B* and *POU5F1*, *LEFTY2* exhibits cell type specfic expresion patterns but the methylation signatures are not clearly cell type specific. The promoter tends to be only slightly more methylated in pluripotent cell lines.

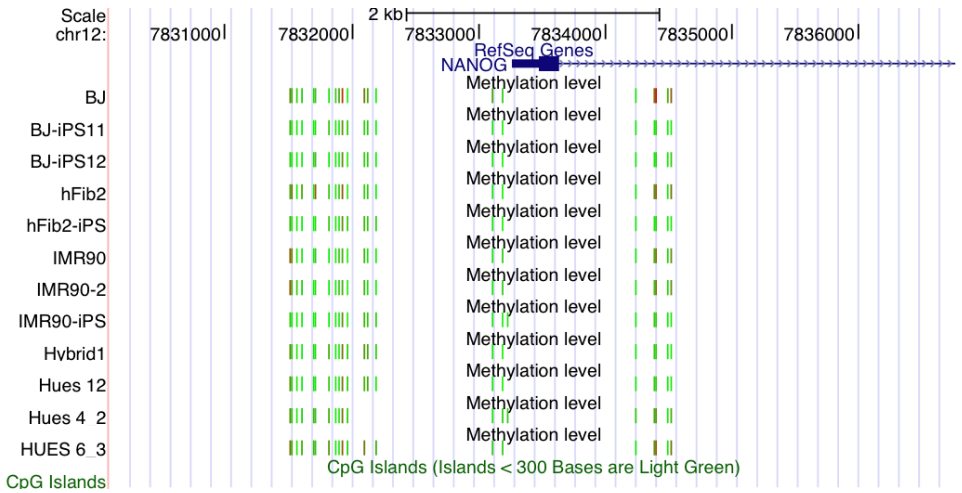


Figure 3.5.6 The *NANOG* TSS region does not show any clear cell type specific methylation patterns though it is highly expressed in pluripotent cell lines only. The regulation of *NANOG* is either DNA methylation independent or occurs at distal regulatory regions.

3.6 DNA Methylation in CpG Islands

Targeting CpG islands across chr12 and chr20, the Cpg30k probe set revealed that the majority of CpGs is unmethylated across pluripotent and fibroblast cell types (Figure 3.6.1). When comparing the CpG islands' methylation states across pluripotent and fibroblast cell types, the majority of CpG islands also does not exhibit significant methylation changes. Figure 3.6.1 shows three clusters of CpG islands where each cluster encompasses (1) strongly correlated, (2) weakly correlated, or (3) non- or anti-correlated CpG islands across pluripotent and fibroblast cell types. Cluster 1 encompasses 76% of all covered CpG islands, which includes most of the unmethylated CpG islands. Cluster 1 contains the remaining 13% of CpG islands and Cluster 3 contains 11% of CpG islands. Figure 3.6.2 shows a portion of the HOXC gene cluster and the methylation states of the contained CpG islands. Most of the CpG islands are within cluster 1 since they show no differential methylation between fibroblasts and pluripotent cell lines but there are also CpG islands that belong to cluster 3. Unlike in the TSS regions, regions showing differential methylation across CpG islands (i.e. cluster 3), when associated with nearby genes, do not show significant gene expression changes.

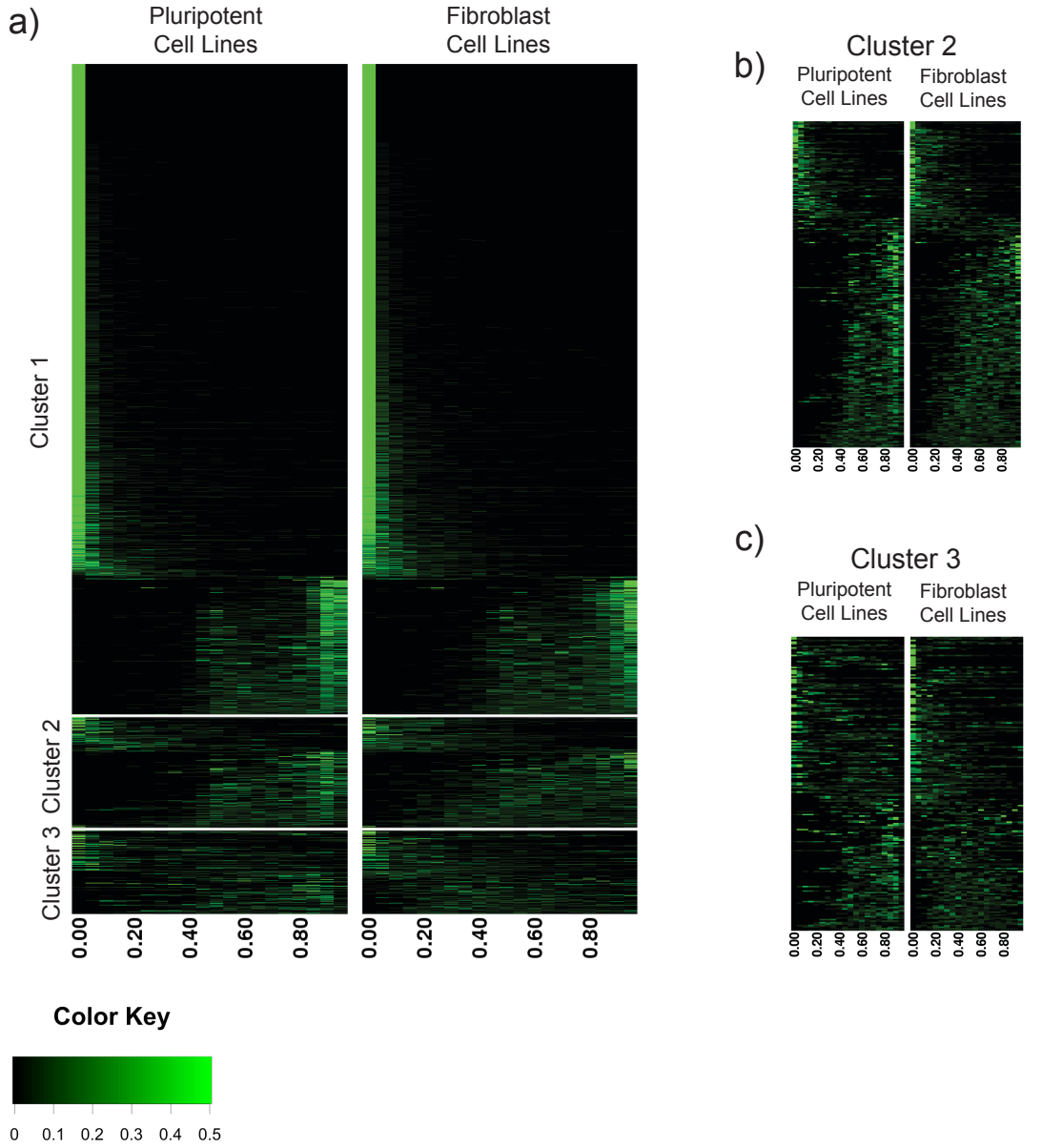


Figure 3.6.1 a) CGIs were clustered into three groups based on the their methylation frequency correlation across cell types. CpG islands that had strong correlation are in cluster 1 (R > .5). CpG islands with weak correlation are in cluster 2 (0 < R< .5). CpG islands showing no or anti-correlation are in cluster 3 (R ≤ 0). Each row represents a CpG island. CpG island labels are omitted for clarity. CpGs are placed into 20 bins based on their methylation frequency values. For example, if a methylation frequency is between .20 and .25, it is placed in bin 4, which is labeled as 0.20. The color of each bin represents the fraction of probes in the TSS region that fall within this bin. b) This is a close-up of Cluster 2 c) This is a close-up of Cluster 3.

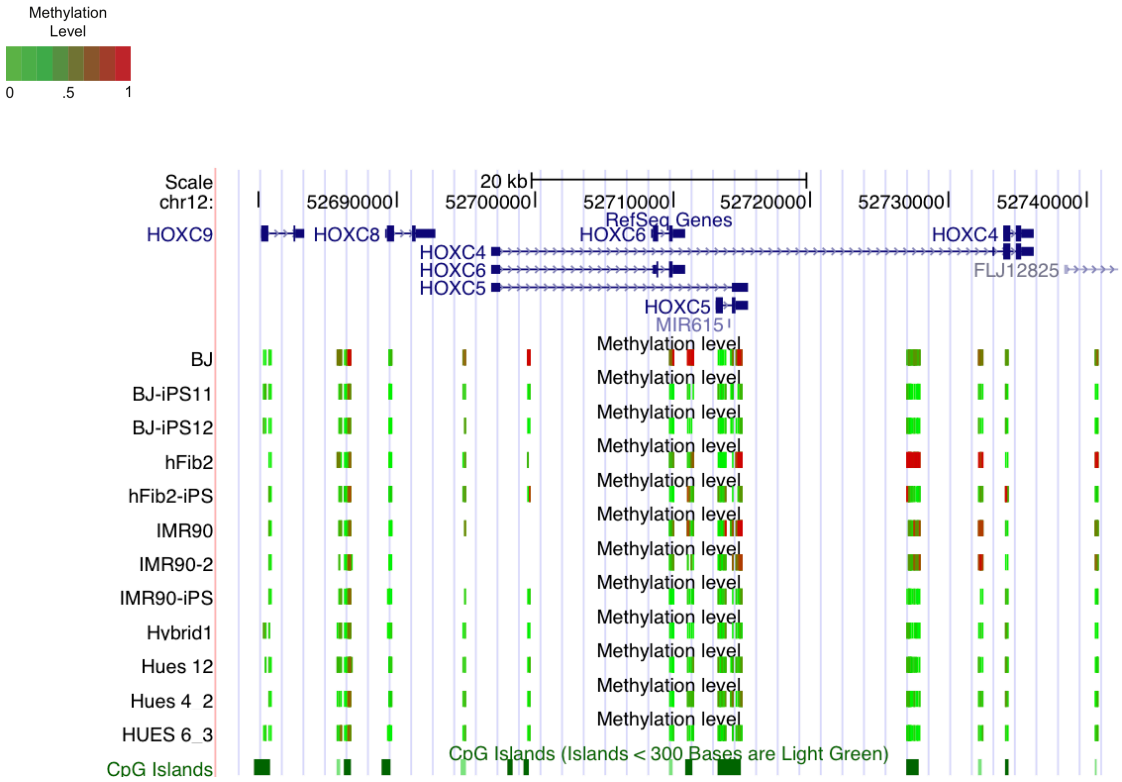


Figure 3.6.2 This figure offers an overview of the CpG islands that cover the HOXC gene cluster. A few CpG islands in this region are located in cluster 3 since they are differentially methylated between fibroblast and pluripotent cell lines. However, most CpG islands show no differential methylation between fibroblasts and pluripotent cell lines (cluster 1). Near the *HOXC9* TSS is a CpG island that is unmethylated across all cell lines (cluster 1 CpG island) and a CpG island near a *HOXC5* variant is strongly methylated only in the BJ cell line.

3.7 Gene Expression and CpG Methylation Levels

The effects of methylation on gene expression vary depending the CpG's position relative to the TSS (Figure 3.7.1). The correlations in the figure were based on a 500 bp window starting at -10 kb and moving in 200 bp increments. CpG methylation values within the window were averaged and compared to the log2 expression value of its associated gene. Significance was based on a background normal distribution created by randomly adding CpGs to a CpG window regardless of the CpGs actual distance from a TSS. CpG windows nearby multiple genes were excluded from this analysis.

Methylation near the TSS tends to decrease gene expression. However, methylation downstream of the TSS correlates with positive gene expression. Methylation 9-10 kb upstream of TSS also exhibits a positive correlation with gene expression. These relationships indicate that DNA methylation, which is targeted mostly at CpG islands, affects transcription differently based on its position relative to the TSS. This correlation analysis included six cell lines and a complete analysis that included additional cell lines, which was published last year, further supports these position effects(*38*). Using the methylome data sets for IMR90 and H1 from Joe Ecker's group, the effects of CpG methylation on gene expression patterns observed in the Cpg30k dataset are confirmed (Figure 3.7.2)(*1*). Methylation is significantly anti-correlated with gene expression around the TSS and methylation is correlated with gene expression away from the TSS. The data

from Joe Ecker's group shows that methylation downstream of the TSS is slightly more correlated with gene expression than upstream methylation. Although the correlations from these analyses are significant given a random background, they are still small values (i.e. less than .5). This demonstrates that other mechanisms are also important in regulating transcription and thusly methylation frequency data cannot be used as a sole source of data to accurately predict gene expression. However, this analysis shows that DNA methylation can affect gene expression positively and negatively depending on its location relative to the TSS of interest.

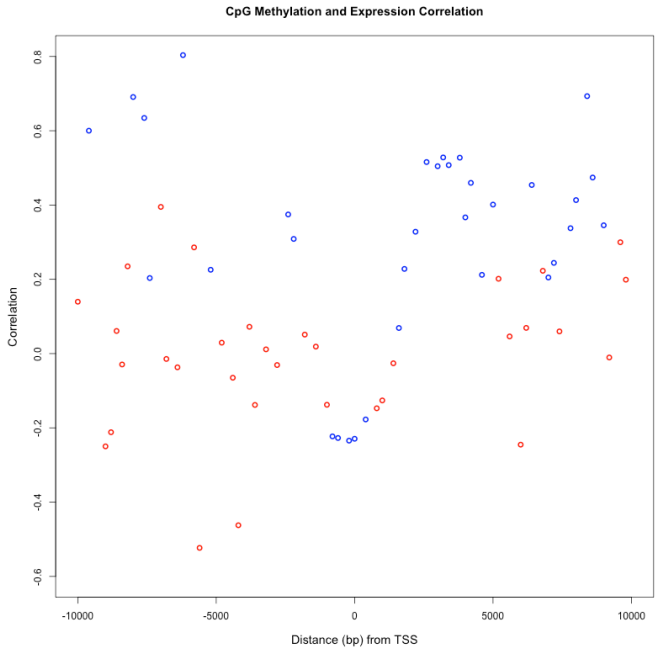


Figure 3.7.1 The figure shows the correlation between CpG methylation values and RefSeq gene expression values across 6 cell lines (i.e. hFib2, hFib2-iPS4, IMR90, IMR90-iPS, BJ, and BJ-iPS12). Significant correlations are in blue (p-value < .01) while non-significant correlations are displayed in red. Methylation correlates with increased expression downstream of the TSS. Around the TSS ([+1 kb, -1kb]), methylation correlates negatively with gene expression. Far upsteam from the TSS, increased methylation correlates with gene expression.

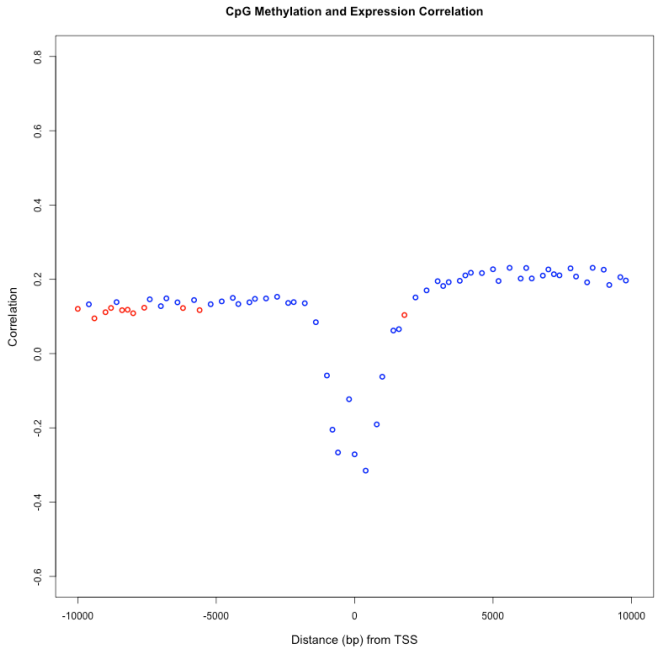


Figure 3.7.2 The figure shows the correlation between BS shotgun sequencing CpG methylation values and RefSeq gene expression values across IMR90 and H1 cell lines(*1*). Significant correlations are in blue (p-value < .01) while non-significant correlations are displayed in red. Methylation correlates with increased expression downstream and upstream of the TSS. Methylation near the TSS correlates with decreased gene expression.

3.8 Conclusion

My analyses revealed that the Cpg30k BSPP was able to capture cell type specific methylation patterns and that methylation correlates with gene transcription. The probe set, which focused on CpG islands across chromosomes 12 and 20, successfully and reliably captured methylation information across 11 cell lines. Clustering these cell lines based on their methylation signatures clearly separated pluripotent cell lines, which included iPS, Hybrid, and ES cell lines, from the differentiated cell lines (i.e. fibroblasts). iPS cell lines appeared to be slightly more methylated than ES cell lines. Both pluripotent cell lines were more methylated than the fibroblast cell lines. Looking at individual TSS regions, DNA methylation signatures do reflect the cell type specific gene expression patterns of several genes. However, many genes expressed in a cell type specific manner did not show cell type specific methylation patterns. This indicates regulation outside of the TSS region and/or DNA methylation independent regulation. CpG methylation correlated negatively with gene expression near the TSS ([-1 kb, +1 kb]). Away from the TSS, methylation is slightly positively correlated with gene expression. However, the weak correlation indicates other factors also regulate gene expression. Looking at the targeted CpG islands specifically, most were unmethylated and did not discriminate between cell types. Based on the lack of methylation specificity for most of the Cpg30k targeted regions, newer

probe sets were designed based on regions experimentally shown to exhibit differential methylation.

Chapter 3 is a rewrite of material as it appears in J. Deng, R. Shoemaker, B. Xie, A. Gore, E. M. LeProust, J. Antosiewicz-Bourget, D. Egli, N. Maherali, I. H. Park, J. Yu, G. Q. Daley, K. Eggan, K. Hochedlinger, J. Thomson, W. Wang, Y. Gao, K. Zhang, Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353 (Apr, 2009). The dissertation author was the primary data analyst of this paper.

Chapter 4 Intracellular Methylation Analyses Using the Cpg30k BSPP

4.1 Abstract

In diploid mammalian genomes, parental alleles can exhibit different methylation patterns (allele-specific DNA methylation, ASM), which have been documented in a small number of cases except for the imprinted regions and X chromosomes in females. We carried out a chromosome-wide survey of ASM across 16 human pluripotent and adult cell lines using Illumina bisulfite sequencing. We applied the principle of linkage disequilibrium (LD) analysis to characterize the correlation of methylation between adjacent CpG sites on single DNA molecules, and also investigated the correlation between CpG methylation and single nucleotide polymorphisms (SNPs). We observed ASM on 23~37% heterozygous SNPs in any given cell line. ASM is often cell-type specific. Furthermore, we found that a significant fraction (38~88%) of ASM regions are dependent on the presence of heterozygous SNPs in CpG dinucleotides that disrupt their methylation potential. This study identified distinct types of ASM across many cell types and suggested a potential role for CpG-SNP in connecting genetic variation with the epigenome.

4.2 Introduction

DNA methylation is an epigenetic marker that plays a direct role in transcriptional regulation. DNA methylation patterns are tissue specific. Embryonic stem cells undergoing differentiation show significant changes in DNA methylation patterns(*1, 38, 41*). In addition to DNA methylation pattern differences between cell lines, DNA methylation can also be allele specific within a cell line and is thusly linked to allele specific gene expression (ASE). An example of allele specific methylation (ASM) is genetic imprinting, which describes the parent-specific gene expression behavior of a small set of genes. The methylation pattern of imprinted genes is distinct; the inactive allele is significantly more methylated than the actively expressed allele. The number of currently known imprinting genes is suspected to be a small fraction of the total number of imprinted genes. Recent work has provided a large candidate list of imprinted genes(*52*) though most of these candidates still remain to be validated. Even less is known is about the genes that fall into the broader ASM category. The biological importance of methylation is clear as disturbances of known methylation patterns are linked to disease phenotypes(*206, 212, 213*).

In a recent survey of DNA methylation changes during nuclear reprogramming of human fibroblasts to induced pluripotent stem cells, 7.6% of CpG islands were found to be dominated by CpG sites that have intermediate levels of methylation (0.25-0.75, referred as fuzzy methylation), even though

the samples used in the assay are polyclonal or monoclonal cell lines(*38*). A small fraction of fuzzily methylated CpG dinucleotides are related to X inactivation and imprinting is unlikely the dominant effect that explains the other regions. ASM is another mechanism that could potentially explain fuzzy methylation. However, only a handful of ASM regions have been identified to date(*214, 215*). Taking advantage of the high-resolution CpG methylation information generated from targeted bisulfite sequencing, we carried out a systematic study to characterize ASM and its role in fuzzy methylation (Figure 4.2.1).

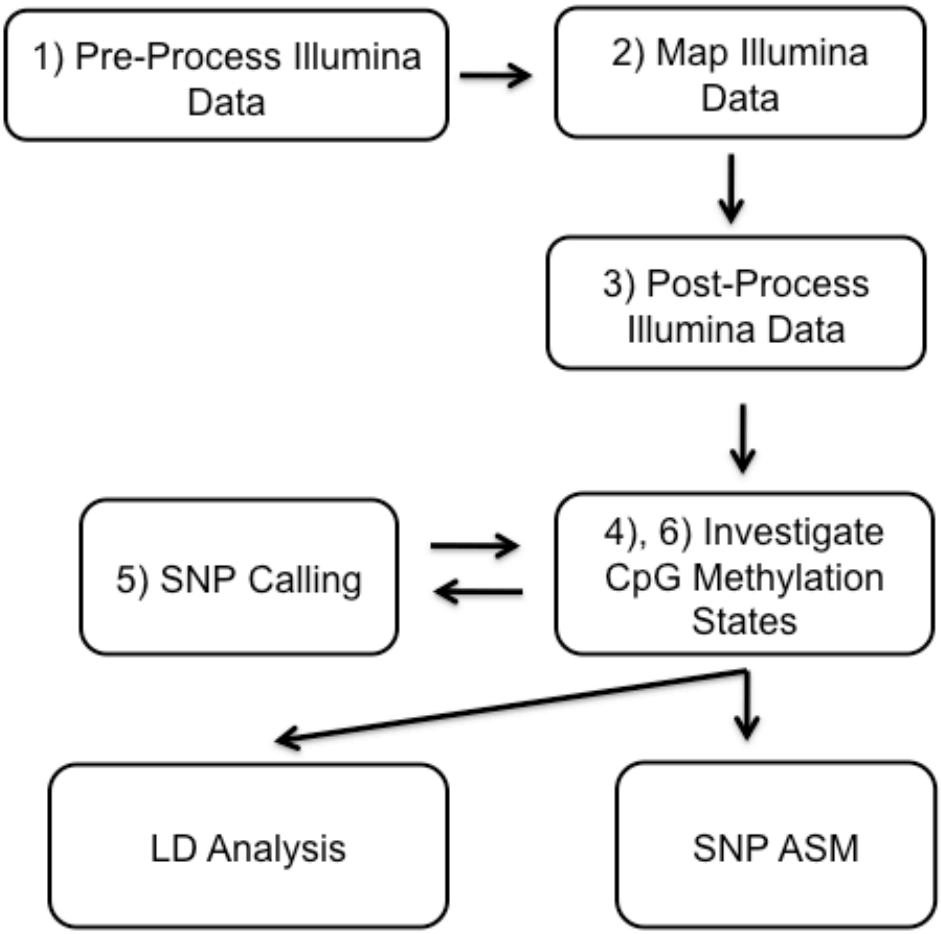


Figure 4.2.1 The workflow of data analysis contains the following steps: 1) Pre-processing of Illumina data involved detecting whether the read was a PCR product or bisulfite converted sequence and taking the reverse complement of PCR product sequences. Then the read sequences were demethylated *in silico*. 2) Reads were mapped to the unmethylated bilsulfite converted hg18 reference genome. 3) Reads that did not uniquely map to the reference are filtered out. Paired end reads that mapped too closely together or too far away were treated as single end reads. 4) The read coordinates of the mapped *in silico* demethylated reads were assigned to the read's original sequence. The methylated state of any known CpGs covered by the reads were investigated. 5) SNPs were called based on the mapped read data using our own algorithm and SAM Tools. The intersection of SNP calls from these two algorithms was used in further anlayses. 6) The methylation states of new CpGs created by SNPs were investigated. 7) The methylation information was used for LD and SNP ASM analyses.

4.3 Results

We reasoned that, if fuzzy methylation were due to unequal methylation levels between the two copies of the chromosomes in the same cells or the presence of heterogeneous epigenetic states among the cell population, the methylation levels on adjacent CpG sites of the same DNA molecules should be highly correlated. When performing Illumina sequencing on bisulfite converted DNA, a sequencing read often contains multiple CpG sites, which can be treated as methylation haplotypes. Such methylation haplotypes are similar to SNP haplotypes, which allowed us to extend the concept of linkage disequilibrium (LD) analysis to characterizing the co-methylation of CpG sites on single DNA molecules. Specifically, we used the LD measurement r^2, which indicates the fraction of variation (of methylation status) on the CpG site A that can be explained by the variation on another CpG site B. Note that in this context LD is defined on a population of diploid cells.

Our analyses were based on the targeted bisulfite sequencing data (41bp reads) previously generated on 11 human pluripotent and adult cell lines(*38*), plus additional paired 36bp Illumina sequencing reads on 8 human cell lines (3 cell lines were covered in both sets). The following regions were examined in this analysis: (1) all 2,020 CpG islands on human chromosomes 12 and 20, (2) 237 promoters in eight ENCODE (the Encyclopedia of DNA Elements) regions, and (3) the 4-kb region centered around the transcription

start sites (TSS) of 26 genes related to development or pluripotency. In addition to the Illumina short reads, Sanger sequencing data from cloned bisulfite PCR amplicons on a selected number of regions were also generated (Tables 4.3.1-4.3.3). Expanding on the previous published read mapping strategy (*38*), we used whole genome mapping for Illumina data and developed methylation haplotype identifying algorithms for Illumina and Sanger sequences. Regions with at least 10x read depth were used in our LD analysis.

Table 4.3.1 SNPs and ASM by Sanger sequencing. ASM statistics of the Sanger sequences compared to the Illumina read regions. Minimum allele read depth shows the minimum number of aligned Sanger sequences for either allele at each SNP region. The ASM categories were determined using the same criteria as for the Illumina sequences. Two called heterozygous SNPs in HUES63 and hFib2 from the Illumina data were not found in Sanger sequences. Illumina and Sanger ASM categories were considered consistent if both sequencing methods called ASM (i.e. category I or II). There were 7 cases where the ASM calls were not consistent between the Illumina and Sanger data. Rs2072788 in PGP1F and PGP1-iPS1 was a SNP that destroyed the cytosine of a CpG dinucleotide in one of the alleles in the Sanger sequences, which made this region ineligible for a category II ASM label. These two SNP regions were considered category I ASM in the Illumina data because of a neighboring CpG site that showed ASM. This corresponding CpG site in the Sanger data did not have sufficient read depth for a statistical call of ASM. Rs20889908 in PGP1F and rs2018002 in PGP3L also had insufficient Sanger read depth for ASM categorization. The discrepancy involving rs1061726 in BJ-iPS11 cannot be explained by read depth or single stranded coverage.

Cell Line	SNP 129 ID	Associated Ref Seq Genes	Sequenced Strand	Minimum Allele Read Depth	Sanger ASM Class	Illumina ASM Class
BJ	rs3746459	NECAB3:intron	-	1	III	III
BJ	rs7266947		+	7	I	I
BJ-iPS11	rs10877897		+	46	I	I
BJ-iPS11	rs1061726		+	6	I	III
BJ-iPS11	rs3746459	NECAB3:intron	-	4	I	I
BJ-iPS12	rs10846023	FLJ22662:intron	-	9	I	I
hFib2	rs2277324	SLC26A10:5 prime	+	0	NA	I
HUES42	rs2277324	SLC26A10:5 prime	+	2	III	III
HUES63	rs220030	SNRPN:intron	-	0	NA	III
IMR90-iPS	rs2018002	DNMT3B:intron	+	2	III	III
IMR90-iPS	rs2089908	LSP1:5 prime	+	8	I	II
PGP1F	rs2236416	MMP9:intron	+	11	III	III
PGP1F	rs2089908	LSP1:5 prime	+	7	III	II
PGP1F	rs2072788	MATN4:exon	-	12	III	I
PGP1L	rs2236416	MMP9:intron	+	6	I	I
PGP1L	rs2072788	MATN4:exon	-	8	I	I
PGP1-iPS1	rs2072788	MATN4:exon	-	11	III	I
PGP3L	rs6061990	TAF4:intron	+	14	II	II
PGP3L	rs2018002	DNMT3B:intron	+	1	III	I
PGP9L	rs6061990	TAF4:intron	+	13	II	I
PGP9L	rs2018002	DNMT3B:intron	+	4	III	III

Table 4.3.2 Primer Sequences.

Index SNP	Forward primer	Reverse primer	Amplicon size (bp)
rs1061726	GTAGAAATTTGGAAGTGGAAATTT	TAATCAATAATTTTTCCAAAAAAAA	645
rs10846023	TTTAGGGGTTGTTAGAGGGTTAGA	AAATTTTTAAAACCAACCCAAACTC	674
rs2018002	GTTTTTGTTTTGGGAAAAGTTAAG	CAAAAAACAACTCAAAATTCATACT	773
rs2072788	GTGAGGTTTTGTTGATTTAGGAGAG	ATCCAAACATTTAAATTAAAAATTC	671
rs2089908	AAGTTTTTGTTGGTTGGATTTTTTA	AAAAAAACCCATATTACCCCTAATC	746
rs2236416	TGGGTTAAAGAATAGGATATATTTGG	AAAAAACCCAAAACCTTAAATAAAC	643
rs3746459	GAAGTTAGGAAATAGTGTGGAGT	AATATACCCAAAACAATAACCC	640
rs6061990	AGGTTTGGGTTATTTTATTTGTTTG	ACTTTCCCAACTCTCAAAACTCTAC	753
rs2277324	GGTTAAGGATGTTTGTAGAAA	ATTAAAACCTCTCACCCTAAAACAC	557
rs220030	TAGGTTGTTTTTTGAGAGAAGTTAT	CTTTTAAAAAAATTTCAAATCTAAC	559
rs10877897	GAGATGATGGTTTGGATTTTTTAG	AAAATCTTTAACCACTACCTACCC	615

Table 4.3.3 Summary of Illumina sequencing data. Single end data is taken from the previously published data set and the paired end data is from new experiments. Cell lines labeled as a mixture of reads contain single end and paired end Illumina sequences, which were merged together before alignment.

Cell Line	Total Reads	Mapped Reads	% Mapped	Single End, Paired End, or Mixture.
BJ	18,516,100	9,319,863	50.33%	Mixture
BJ-iPS11	12,276,017	5,235,162	42.65%	Single
BJ-iPS12	10,539,126	4,587,661	43.53%	Single
hFib2	14,408,007	5,780,509	40.12%	Single
hFib2-iPS	14,805,000	5,652,624	38.18%	Single
HUES12	18,764,749	8,849,367	47.16%	Mixture
HUES42	11,185,368	6,378,978	57.03%	Single
HUES63	14,315,301	5,811,519	40.60%	Single
Hybrid1	16,522,597	9,093,421	55.04%	Mixture
IMR90	24,687,802	10,493,117	42.50%	Single
IMR90-iPS	10,985,469	5,445,173	49.57%	Single
PGP1F	8,441,376	5,021,880	59.49%	Paired
PGP1-iPS1	5,617,832	3,611,457	64.29%	Paired
PGP1L	5,258,188	3,704,072	70.44%	Paired
PGP3L	5,286,442	3,481,123	65.85%	Paired
PGP9L	5,689,552	4,121,764	72.44%	Paired

We first validated our LD analysis on known imprinted regions and female X chromosomes. In such regions we expected to observe high r^2 values that extended over a long distance. We developed an algorithm to search for such regions (see Methods). Similar to the LD blocks in the human population, we called these regions “methylation LD blocks”. A block was first created between a pair of CpG sites with $r^2 > 0.3$ and extended if another CpG pair with $r^2 > 0.3$ was less than 100 bp away. Since we were most interested in examining extended organized methylation regions, we filtered out LD blocks that spanned less than 100 bp or contained less than 10 CpG pairs with $r^2 > 0.3$. We observed LD blocks in all 16 cell lines for the known imprinted genes *SNRPN* and *GNAS*(*52*). *NNAT*, another known imprinted gene, was found to contain LD blocks in 13 cell lines. There were two other imprinted genes found but the read coverage for these genes was much lower than for *SNRPN*, *GNAS*, and *NNAT*. *IGF2AS* was only covered in three cell lines and *NDN* was only covered in one cell line. No LD blocks were found within these two gene regions in their respective cell lines. In female cell lines, approximately 43~75% of the X chromosome regions included in our analysis were covered by LD blocks which clearly distinguished them from the majority of male lines due to X-chromosome inactivation. A much lower fraction (0~9%) of X chromosome regions was within LD blocks for the male cell lines. The only exception is HUES63, which is a male human embryonic stem cell line, exhibited extended LD on the X chromosome. This is likely due to the presence of a subpopulation of cells that were already in the early stage of

differentiation (Figure 4.3.1). Bisulfite Sanger sequencing of a known imprinted gene, *SNRPN*, from HUES63 revealed a methylation LD block that extended over 500bp. In such regions, LD does not decay over the physical distance (Figure 4.3.2a). However, imprinted regions are not free of noise. We also observed many pairs of CpG sites that have little or no correlation (Figure 4.3.2b-c).

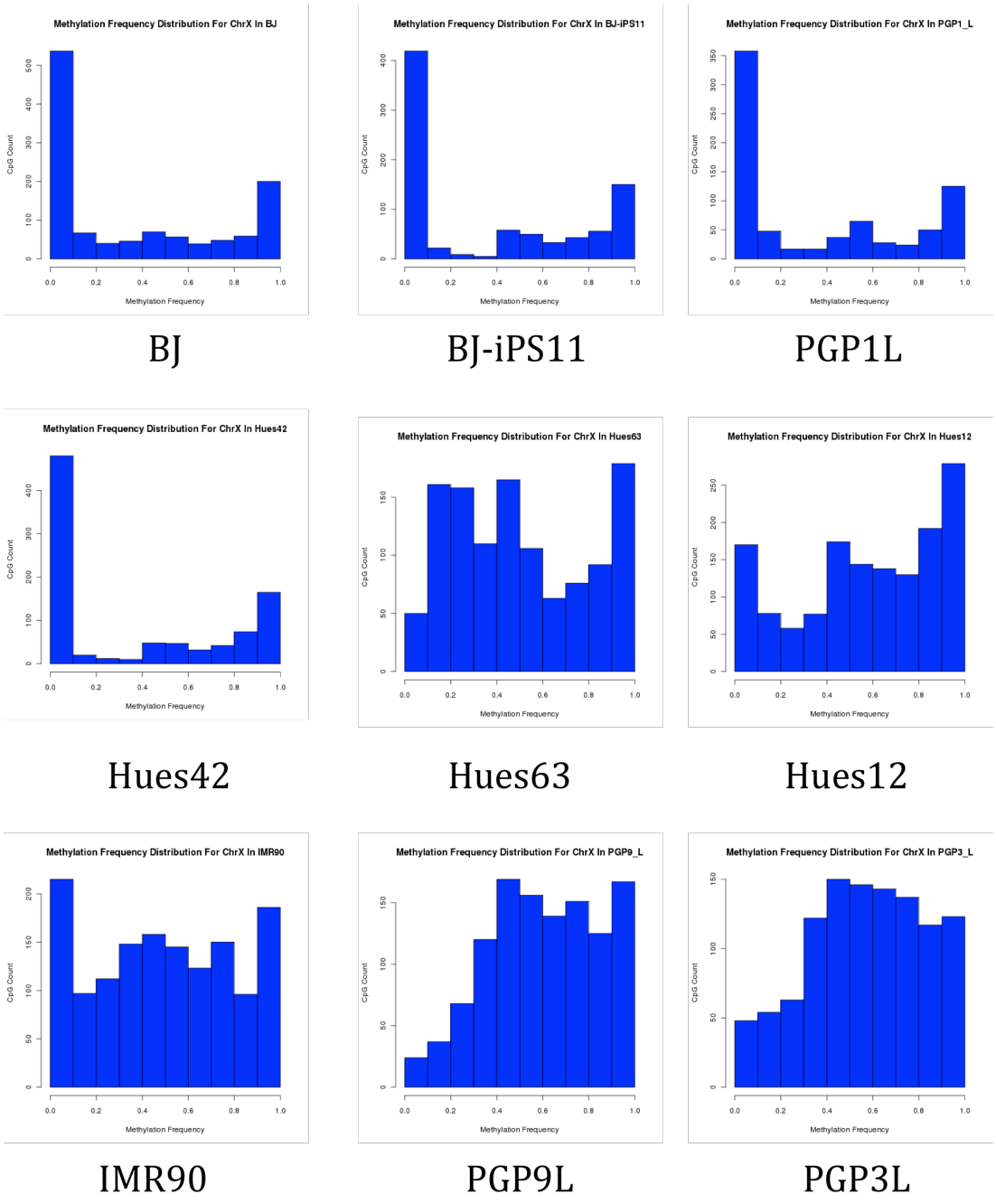


Figure 4.3.1 Methylation frequency histograms for CpG dinucleotides located in chromosome X. The X-axis represents the methylation frequency bins and the y-axis represents the number of CpG dinucleotides within a certain methylation frequency bin. A methylation frequency value of one represents complete methylation at a CpG. HUES63, a male cell line, has a methylation frequency distribution in chromosome X that visibly contains more fuzzily methylated probes (.25 < methylation frequency <.75) than other male cell lines. Fuzzily methylated probes in chromosome X are indicative of female cell lines (HUES12, IMR90, PGP3L, PGP9L) rather than male cell lines (BJ family, HUES42, PGP1L), which is due to allele specific methylation.

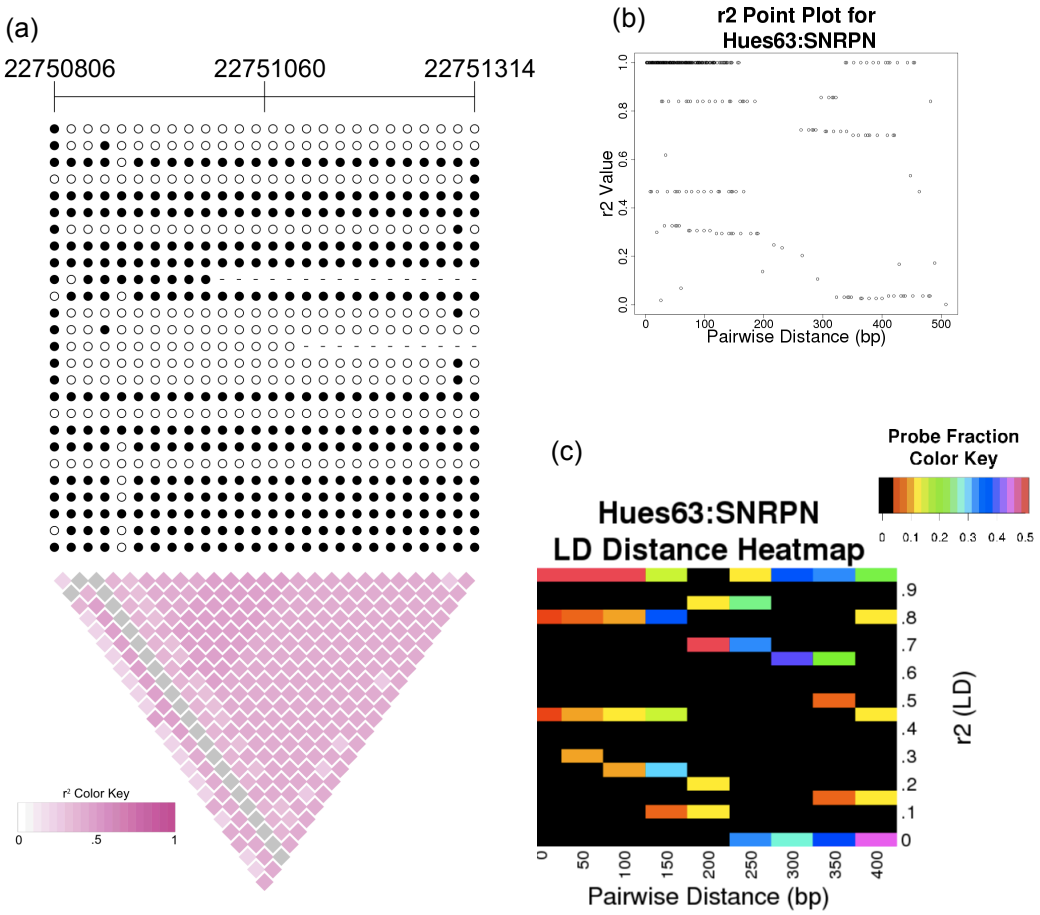


Figure 4.3.2 Linkage disequilibrium (LD) analysis of CpG methylation haplotypes. (a) LD diagram of 5' prime region of the imprinted gene *SNRPN* (chr15:22750806-22751314) from HUES63. Each row represents a Sanger sequence and each column represents a CpG dinucleotide. Filled circles indicate methylated CpGs, empty circles indicate unmethylated CpGs, and dashed lines mean the methylation state of the CpG could not be determined. Chromosomal coordinates are listed above. This region shows a methylation LD block spanned over 500bp in Sanger reads. (b) and (c) CpG pairwise r^2 value plots and heatmap for 5' prime region of SNRPN gene. While the majority CpG pairs have high methylation correlation values ($r^2 > 0.3$), some pairs of CpG sites have little or no correlation ($r^2<0.3$) The pairwise distance represents the separation of the CpG dinucleotides used in the r^2 calculation. The heatmap colors represent the probe fraction at a given pairwise distance (rounded down to the nearest 50 bp) that has the indicated r^2 value. The probe fractions for each pairwise distance sum to one. The color scale saturates at 0.5 so that small probe fraction differences can be distinguished

We next extended the LD analysis to the other autosomal regions. Various levels of LD were observed in all 16 cell lines included in this study (Figure 4.3.3). Pluripotent cell lines appear to exhibit more extended LD compared with the corresponding fibroblast lines, which can potentially be explained by the less compact chromatin structure in pluripotent. The more open chromatin structure would allow cis-regulation to operate over a longer distance(*216*). Of the 108-289Kb genomic regions with sufficient read coverage for the LD analysis, 10~24% were within methylation LD blocks, which vary in both the number of CpG sites per block and the block size (Table 4.3.4). The majority of the methylation LD blocks (68~94%) are not known to be involved in genomic imprinting or X inactivation.

Many genes were found to contain LD blocks across several cell lines. Out of 309 genes that are associated with at least one LD block in at least one cell line, 100 genes (32%) contain LD blocks in at least 5 cell lines. Eighteen genes (5.9%) are found to have LD blocks in at least 12 cell lines (Table 4.3.5). *GNAS* and *SNRPN* contain LD blocks in all 16 cell lines and *NNAT* contains LD blocks in 13 cell lines. Overall 31~50% of fuzzily methylated CpGs had strong LD (r^2 >0.3) with at least one adjacent site, and 14~36% of fuzzily methylated regions were found within methylation LD blocks. A fraction of fuzzily methylated CpGs has neighbors with correlated methylation patterns but many of these patterns are too localized to meet our definition for a methylation LD block. The organized CpG methylation patterns within LD blocks cannot be explained by stochastic effects. They are evidence of either

epigenetic heterogeneity across cells or preferential methylation of one particular parental chromosome copy.

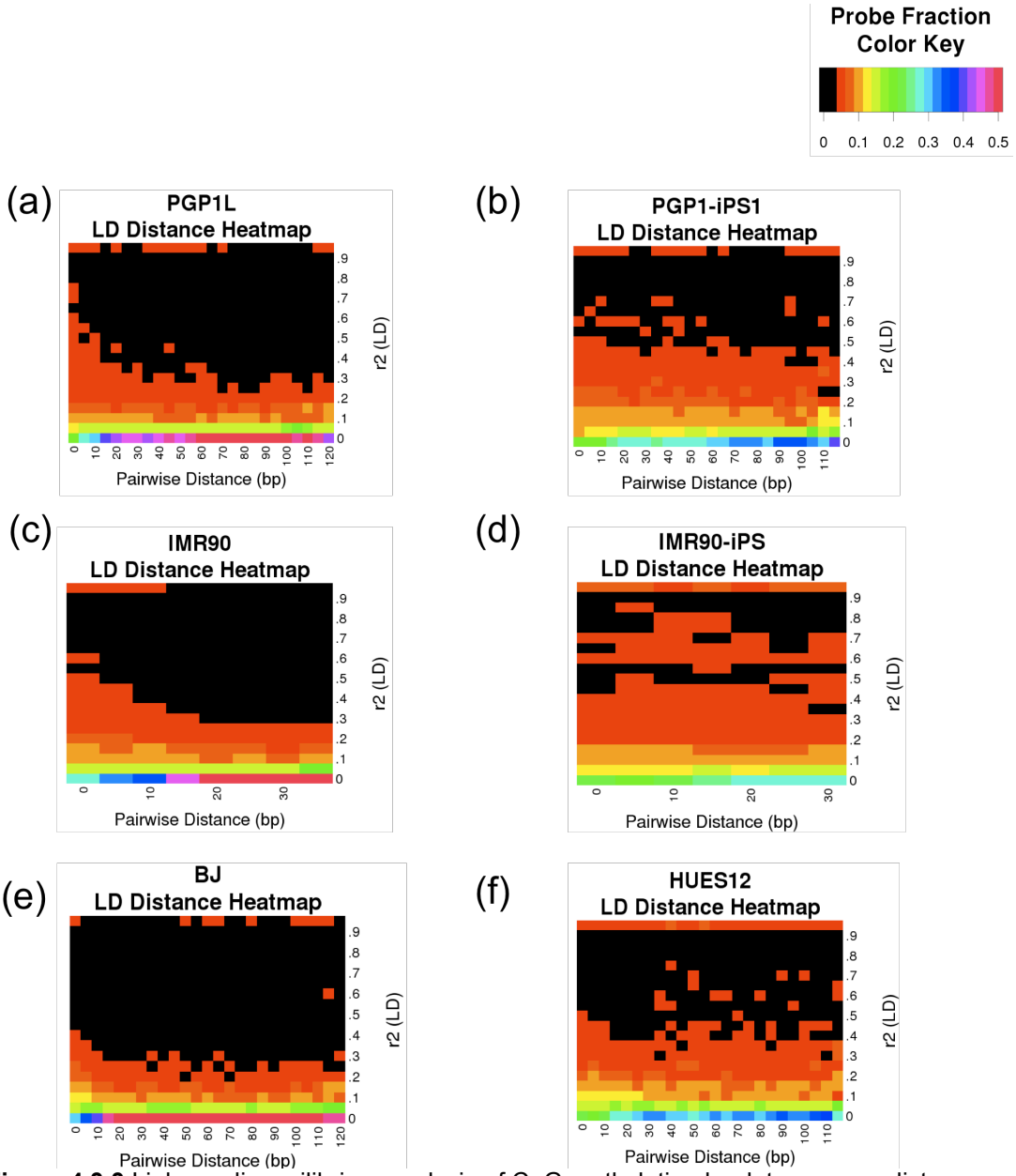


Figure 4.3.3 Linkage disequilibrium analysis of CpG methylation haplotypes over distance (bp). (a),(c), and (e) are from differentiated cells and the heatmap shows that the r^2 value distribution shifts downwards as pairwise distance increases. (b), (d), and (f) represent undifferentiated cells and the distribution of r^2 values is maintained as pairwise distance increases. These figures show that undifferentiated cells contain organized methylation patterns that tend to span farther distances than in differentiated cells.

Table 4.3.4 LD Coverage Statistics. LD blocks are defined as regions that span more than 100 bp and contain more than 10 CpG dinucleotide pairs whose $r^2 > 0.3$. Female cell lines, containing two X chromosomes, are expected to show LD blocks in regions of X-chromosome inactivation. The rightmost column shows the percentage of LD blocks not located in chromosome X or known imprinting genes. Total regions analyzed include the base pairs between CpGs that had sufficient read coverage for the LD analysis.

Cell Line (Gender)	Total Regions Analyzed (bp)	% LD Block Covered	Chr. X Regions Analyzer (bp)	ChrX % LD Block Covered	% LD blocks not in Chr. X or known imprinted genes
BJ (M)	288,574	11.99%	4,560	0%	84.03%
BJ-iPS11 (M)	117,147	21.73%	747	0%	87.46%
BJ-iPS12 (M)	118,334	22.36%	986	0%	85.76%
hFib2 (M)	119,085	16.84%	1,411	0%	85.96%
hFib2-iPS (M)	129,883	13.52%	1,172	0%	86.32%
HUES12 (F)	246,680	17.17%	9,006	57%	79.20%
HUES42 (M)	126,481	15.92%	929	0%	85.38%
HUES63 (M)	152,400	19.47%	5,616	71.08%	71.09%
Hybrid1 (M/F)	246,159	15.75%	6,406	47.35%	81.94%
IMR90 (F)	200,953	10.35%	7,815	42.47%	68.02%
IMR90-iPS (F)	107,727	21.75%	4,428	74.82%	73.71%
PGP1F (M)	229,597	15.63%	4,394	2.62%	87.18%
PGP1-iPS1 (M)	171,338	24.21%	2,128	8.74%	93.47%
PGP1L (M)	219,643	16.09%	2,386	8.17%	91.11%
PGP3L (F)	245,052	21.22%	10,634	38.83%	85.17%
PGP9L (F)	247,465	21.24%	10,924	47.11%	81.56%

Table 4.3.5 Genes with conserved LD blocks across 12 or more cell lines. These genes contain CpG dinucleotides that are strong candidates for biological regulation via methylation. Cell line names are abbreviated so that they can be printed on a single line.

Gene	Cell Lines	Cell Lines
FRG1B	16	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,P1L,BJiPS11,hFiPS
HM13	16	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,P1L,BJiPS11,hFiPS
NR_003579	16	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,P1L,BJiPS11,hFiPS
GNAS	16	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,P1L,BJiPS11,hFiPS
SNRPN	16	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,P1L,BJiPS11,hFiPS
KCNS1	14	P1F,P1iPS11,IMR90iPS,H12,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,P1L,BJiPS11,hFiPS
KCNK15	13	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,Hy,BJiPS12,P3L,hFib2,hFiPS,P1L
NR_003531	13	P1F,P1iPS11,IMR90,H12,BJ,P9L,H63,H42,Hy,P3L,hFib2,hFiPS,P1L
LOC100134868	13	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,BJiPS11
MEG3	13	P1F,P1iPS11,IMR90,H12,BJ,P9L,H63,H42,Hy,P3L,hFib2,hFiPS,P1L
NNAT	13	P1F,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,hFiPS
NR_004846	13	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,H42,Hy,BJiPS12,P3L,BJiPS11
JPH2	12	P1F,P1iPS11,IMR90,H12,BJ,P9L,H42,Hy,P3L,hFib2,hFiPS,P1L
C20orf96	12	P1F,P1iPS11,H12,BJ,P9L,H63,H42,BJiPS12,P3L,hFib2,BJiPS11,P1L
TXNRD1	12	P1F,P1iPS11,IMR90,IMR90iPS,H12,BJ,P9L,H63,BJiPS12,P3L,BJiPS11,P1L
THBD	12	P1F,P1iPS11,IMR90,IMR90iPS,H12,P9L,H63,Hy,hFib2,BJiPS11,hFiPS,P1L
KRT86	12	P1F,P1iPS11,IMR90iPS,H12,BJ,P9L,Hy,BJiPS12,hFib2,BJiPS11,hFiPS,P1L
KCNQ2	12	P1F,P1iPS11,IMR90,BJ,P9L,Hy,BJiPS12,P3L,hFib2,BJiPS11,hFiPS,P1L
CACNA2D4	12	P1iPS11,IMR90,IMR90iPS,H12,P9L,H63,H42,Hy,P3L,hFib2,hFiPS,P1L
TBX3	12	P1F,P1iPS11,H12,P9L,H63,H42,Hy,BJiPS12,P3L,hFib2,BJiPS11,P1L

To explore whether the methylation LD blocks are due to heterogeneity of cell populations or unequal methylation of chromosome copies, we developed an algorithm for SNP calling from bisulfite sequencing reads (see Methods). Heterozygous SNPs allowed us to distinguish methylation haplotypes from the two parents, and to detect allele-specific methylation. If fuzzy methylation were simply due to the presence of heterogeneity in cell populations, we would not expect to observe allelic preference within a cell. Due to the reduced sequence complexity of the bisulfite converted genome, we also adapted SAMtools to make SNP calls(*217*). We compiled a list of confident SNP calls based on the intersection between our own algorithm and SAMtools. Totally we identified 240-457 heterozygous SNPs and197-391 homozygous SNPs in each cell line (Table 4.3.6). The three cell lines PGP1L, PGP1F and PGP1-iPS were derived from the same individual (PGP1) in the Personal Genome Project, whose full genome was recently sequenced (*218*). We observed a high concordance between the SNPs called from our bisulfite sequencing data and those generated by whole genome sequencing. For the PGP1 lines, 96-97% of SNP calls made by Complete Genomics matched our SNP calls. As expected, cell lines of identical genetic background showed an expected higher number of pair-wise overlapped SNP calls relative to the other cell lines (Tables 4.3.7-4.3.8). From the bisulfite sequencing reads that contain both heterozygous SNPs and at least one CpG site, we identified methylation that significantly associates with one allele of a SNP using the Fisher's Exact Test (see Methods). We also required a minimum methylation

frequency difference of 0.1 between the alleles for ASM categorization. In each cell line, 23~37% of heterozygous SNPs were found to associate with ASM (Table 4.3.9). Only a small fraction of ASM (6%) was consistent across all cell lines in which the heterozygous SNPs are present. The remaining cases are either cell type specific or individual specific (Figure 4.3.6). We also compared the allelic methylation frequencies of 980 individual CpG sites that are linked to SNPs from two batches of IMR90 fibroblast cultures. Excellent correlation was observed between the biological replicates (Pearson correlation coefficient $r^2 = 0.90$), which indicated that our observations were not due to technical artifacts or biological fluctuations.

Table 4.3.6 SNP calling statistics. Candidate sites include all chromosomal locations with sufficient read coverage (minimum 10x coverage on analyzed strands) and sequence quality. Double stranded sites were checked for reverse complementarity. SNP calls were filtered so that only SNP calls made at rs129 sites that matched the reported SNP bases were recorded.

Cell Line	Total Candidate SNP Sites Examined	Called Heterozygous SNPs	Heterozygous SNP Call Percentage	Called Homozygous SNPs	Homozygous SNP Call Percentage
BJ	1,525,138	457	0.030%	391	0.026%
BJ-iPS11	1,123,678	381	0.034%	260	0.023%
BJ-iPS12	1,192,417	402	0.034%	273	0.023%
hFib2	1,350,920	279	0.021%	307	0.023%
hFib2-iPS	1,144,060	283	0.025%	233	0.020%
HUES12	1,348,641	391	0.029%	344	0.026%
HUES42	1,128,168	308	0.027%	283	0.025%
HUES63	1,178,908	382	0.032%	197	0.017%
Hybrid1	1,390,831	395	0.028%	199	0.014%
IMR90	1,434,099	436	0.030%	334	0.023%
IMR90-iPS	1,139,705	430	0.038%	260	0.023%
PGP1F	1,020,601	292	0.029%	253	0.025%
PGP1-iPS1	835,010	240	0.029%	204	0.024%
PGP1L	913,151	257	0.028%	227	0.025%
PGP3L	951,308	254	0.027%	240	0.025%
PGP9L	982,170	272	0.028%	236	0.024%

Table 4.3.7 Pairwise common heterozygous SNPs between cell lines. Each matrix entry represents the number of heterozygous SNPs called in the row i cell line that match with those called in the column j cell line.

Cell Line	BJ	BJ-iPS11	BJ-iPS12	hFib2	hFib2-iPS	HUES12	HUES42	HUES63	Hybrid1	IMR90	IMR90-iPS	PGP1F	PGP1-iPS1	PGP1L	PGP3L	PGP9L
BJ	502	369	378	114	111	152	102	144	219	149	150	113	94	97	104	120
BJ-iPS11	369	421	366	115	116	140	95	136	229	128	137	99	84	85	97	117
BJ-iPS12	378	366	450	111	117	152	104	138	230	141	154	100	95	89	96	124
hFib2	114	115	111	305	241	116	81	98	127	121	122	89	74	70	85	87
hFib2-iPS	111	116	117	241	296	119	79	94	126	120	119	81	73	68	80	82
HUES12	152	140	152	116	119	427	110	119	137	138	123	105	90	96	99	98
HUES42	102	95	104	81	79	110	341	101	116	100	99	97	85	84	79	93
HUES63	144	136	138	98	94	119	101	417	129	134	134	107	86	94	92	85
Hybrid1	219	229	230	127	126	137	116	129	412	135	147	120	104	109	120	125
IMR90	149	128	141	121	120	138	100	134	135	490	360	111	101	102	101	101
IMR90-iPS	150	137	154	122	119	123	99	134	147	360	493	109	97	97	100	105
PGP1F	113	99	100	89	81	105	97	107	120	111	109	322	222	231	85	82
PGP1-iPS1	94	84	95	74	73	90	85	86	104	101	97	222	266	223	71	73
PGP1L	97	85	89	70	68	96	84	94	109	102	97	231	223	290	79	80
PGP3L	104	97	96	85	80	99	79	92	120	101	100	85	71	79	272	89
PGP9L	120	117	124	87	82	98	93	85	125	101	105	82	73	80	89	288

Figure 4.3.8 Pairwise common homozygous SNPs between cell lines. Each matrix entry represents the number of homozygous SNPs called in the row i cell line that match with those called in the column j cell line.

Cell Line	BJ	BJ-iPS11	BJ-iPS12	hFib2	hFib2-iPS	HUES12	HUES42	HUES63	Hybrid1	IMR90	IMR90-iPS	PGP1F	PGP1-iPS1	PGP1L	PGP3L	PGP9L
BJ	391	248	259	121	119	159	122	137	165	155	141	143	121	129	130	132
BJ-iPS11	248	260	229	106	107	129	105	118	140	114	123	124	103	108	114	114
BJ-iPS12	259	229	273	104	105	140	112	118	137	120	125	120	104	115	114	117
hFib2	121	106	104	307	170	127	114	95	101	114	92	120	108	110	125	115
hFib2-iPS	119	107	105	170	233	119	111	87	97	105	98	114	104	105	115	111
HUES12	159	129	140	127	119	344	155	111	102	137	129	145	122	139	140	143
HUES42	122	105	112	114	111	155	283	101	93	121	111	147	121	128	119	123
HUES63	137	118	118	95	87	111	101	197	106	117	116	108	88	92	104	99
Hybrid1	165	140	137	101	97	102	93	106	199	116	109	101	94	87	94	93
IMR90	155	114	120	114	105	137	121	117	116	334	208	128	117	113	116	129
IMR90-iPS	141	123	125	92	98	129	111	116	109	208	260	118	104	105	100	118
PGP1F	143	124	120	120	114	145	147	108	101	128	118	253	183	186	121	129
PGP1-iPS1	121	103	104	108	104	122	121	88	94	117	104	183	204	173	113	114
PGP1L	129	108	115	110	105	139	128	92	87	113	105	186	173	227	114	114
PGP3L	130	114	114	125	115	140	119	104	94	116	100	121	113	114	240	127
PGP9L	132	114	117	115	111	143	123	99	93	129	118	129	114	114	127	236

Table 4.3.9 SNPs associated with ASM. Category I SNPs include SNPs whose ASM is not based solely on SNP containing CpG sites. However, many category I SNP regions still have SNP containing CpG nucleotides. Category II SNPs include SNPs whose ASM is completely dependent on the presence of a SNP containing CpG dinucleotide. Category III SNPs (not shown) do not show ASM. The category I and category II percentages represent the number of SNPs in the respective category divided by the total number of heterozygous SNP calls in each cell line.

Cell line	Het. SNPs called	# ASM regions	# ASM regions with CpG-SNPs	% of Category I ASM	% of Category I ASM containing CpG-SNPs	% of Category II ASM
BJ	457	123	102	11%	59%	16%
BJ-iPS11	381	102	92	5%	70%	22%
BJ-iPS12	402	108	96	7%	61%	20%
hFib2	279	87	58	19%	46%	12%
hFib2-iPS	283	67	58	7%	53%	17%
HUES12	391	114	97	7%	45%	22%
HUES42	308	76	63	7%	52%	17%
HUES63	382	89	76	5%	55%	18%
Hybrid1	395	120	107	8%	67%	22%
IMR90	436	162	128	22%	67%	15%
IMR90-iPS	430	129	120	3%	60%	27%
PGP1_F	292	83	66	12%	59%	17%
PGP1-iPS1	240	62	56	9%	77%	17%
PGP1L	257	88	61	20%	59%	14%
PGP3L	254	94	62	20%	51%	17%
PGP9L	272	74	61	8%	57%	19%

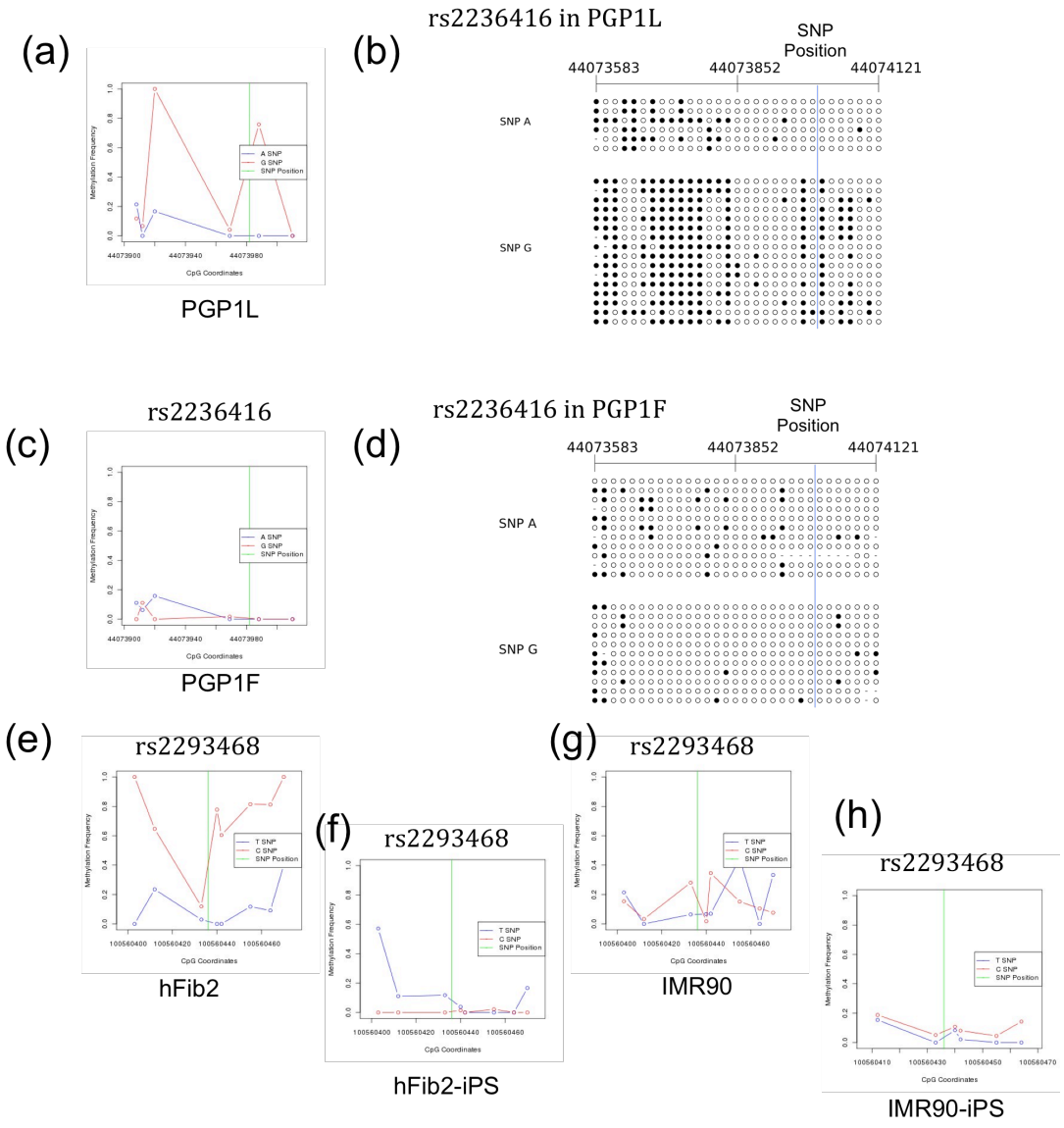
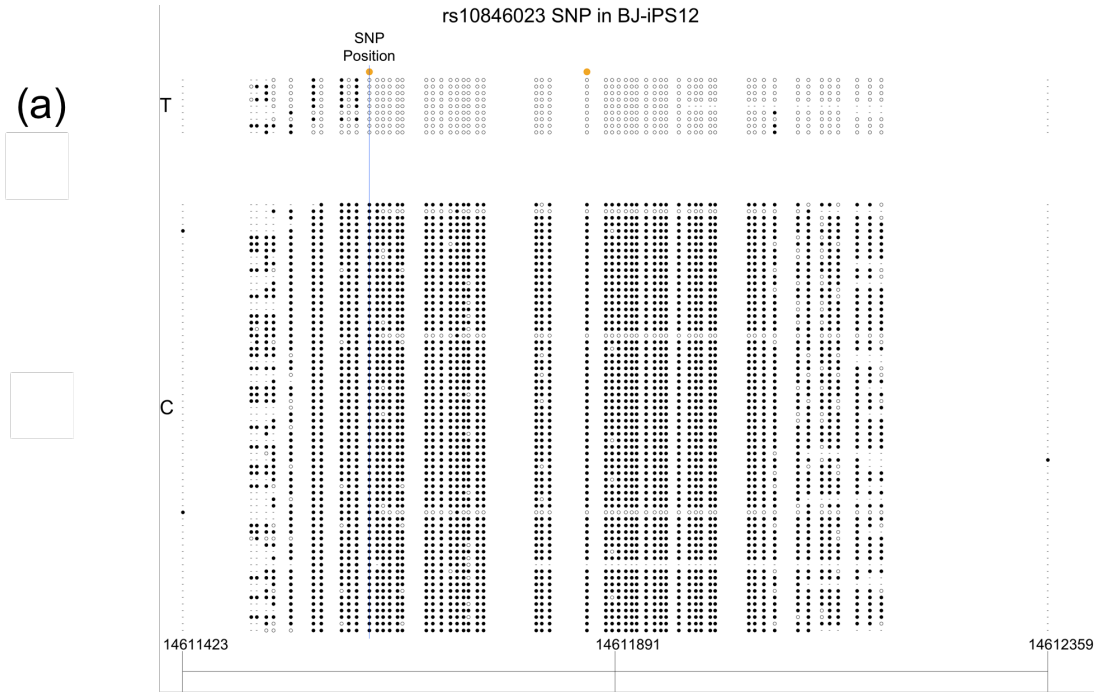
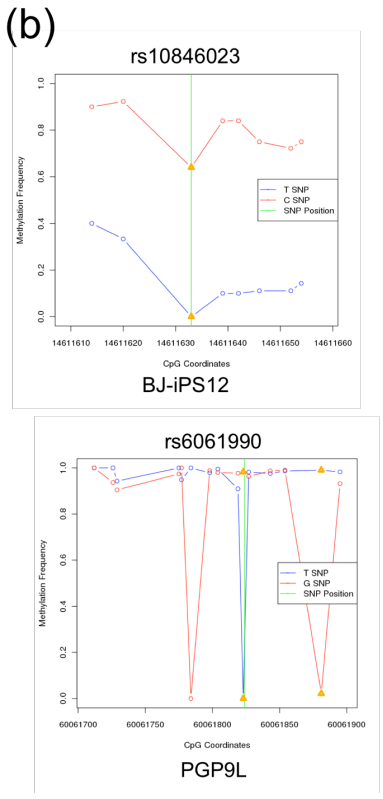


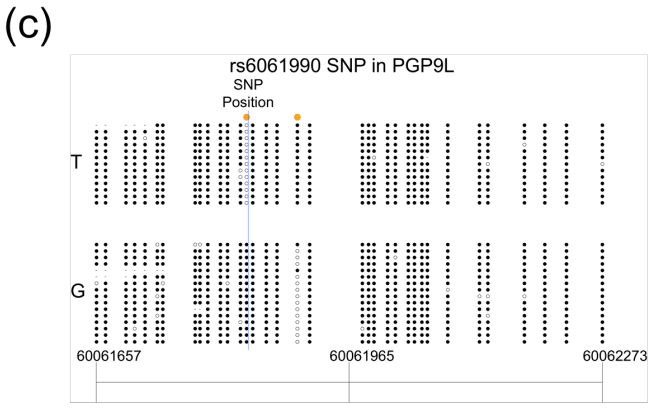
Figure 4.3.6 Cell type specificity and individual dependence of ASM. (a)-(d) Illumina and Sanger sequences from PGP1L and PGP1F showing cell type specificity of ASM in rs2236416 indexed region in the intron of MMP9 gene (chr20:44073908-44074010 in Illumina and chr20:44073583-44074121 in Sanger data). ASM is seen in PGP1L near rs2236416 but not seen in PGP1F. (e)-(h) Illumina reads from hFib2, IMR90, hFib2-iPS (chr12:100560403-100560470) and IMR90-iPS (chr12:100560412-100560464) show the cell type specificity of ASM in rs2293468 indexed region. ASM around SNP rs2293468 is seen in hFib2, hFib2-iPS, and IMR90 but not IMR90-iPS.

We validated 12 SNP sites by performing bisulfite PCR, cloning and Sanger sequencing on one or more cell lines in which SNPs were called. A total of 21 SNP regions were amplified and Sanger sequenced. The Sanger regions that show ASM fall into two categories. In Category I, more than one adjacent CpG site exhibit consistent bias in methylation. The known imprinted regions fall into this category. However, even in autosomal regions not known to be related to genetic imprinting, similar allelic preference can extend over 900bp (Figures 4.3.7 and 4.3.8). In Category II, ASM is highly localized and restricted to only a very small number of CpG sites in a region (Figures 4.3.8 and 4.3.9). Seven sequences had inconsistent ASM classification between the Illumina and Sanger sequencing. Four of these inconsistencies were due to the inability to establish statistical ASM significance due to the low Sanger sequencing read depth. Two SNP sites predicted by the Illumina data were not found in the Sanger data (Table 4.3.1), which are likely due to incorrect mapping of short bisulfite sequencing reads. One SNP site had inconsistent ASM behavior between Sanger and Illumina data. Overall we found a Pearson correlation coefficient of r^2=.78 for the allelic methylation frequencies of CpGs shared between the Sanger sequencing and Illumina data sets.

Figure 4.3.7 Allele specific methylation. (a) Sanger reads from BJ-iPS12 show an extended ASM region in dbSNP 129 rs10846023 indexed region (chr12:14611423-14612359) in the intron of FLJ22622 gene. This ASM region includes two CpG dinucleotides that overlap with SNPs but the ASM is not limited to these sites. Orange circles in the Sanger sequence diagrams represent SNPs that overlap with CpG dinucleotides. (b) An allele specific methylation frequency graph based on aligned Illumina data showing ASM at rs10846023 in BJ-iPS12 (chr12:14611616-14611654). The y-axis represents the methylation frequency where a value of 1 indicates complete methylation at a CpG dinucleotide. The x-axis represents the chromosomal coordinates. (c) Sanger sequence data around SNP site rs6061990 (TAF4 intronic region, chr20:60061657-60062273) in PGP9L illustrates an example where ASM is solely dependent on SNPs at CpG dinucleotides. (d) An allele specific methylation frequency graph based on Illumina data showing ASM at rs6061990 in PGP9L (chr20:60061712-60061895). The green line indicates the SNP position and the orange triangles indicate SNPs that overlap with CpG dinucleotides. Note that the Illumina data shows ASM at a CpG (chr20: 60061784) that is not supported by the Sanger data.


(a)
rs10846023 SNP in BJ-iPS12
SNP
Position
T
C
14611423
14611891
14612359


(b)
rs10846023
Methylation Frequency
T SNP
C SNP
SNP Position
14611610
14611620
14611630
14611640
14611650
14611660
CpG Coordinates
BJ-iPS12
rs6061990
Methylation Frequency
T SNP
G SNP
SNP Position
60061700
60061750
60061800
60061850
60061900
CpG Coordinates
PGP9L


(c)
rs6061990 SNP in PGP9L
SNP
Position
T
G
60061657
60061965
60062273

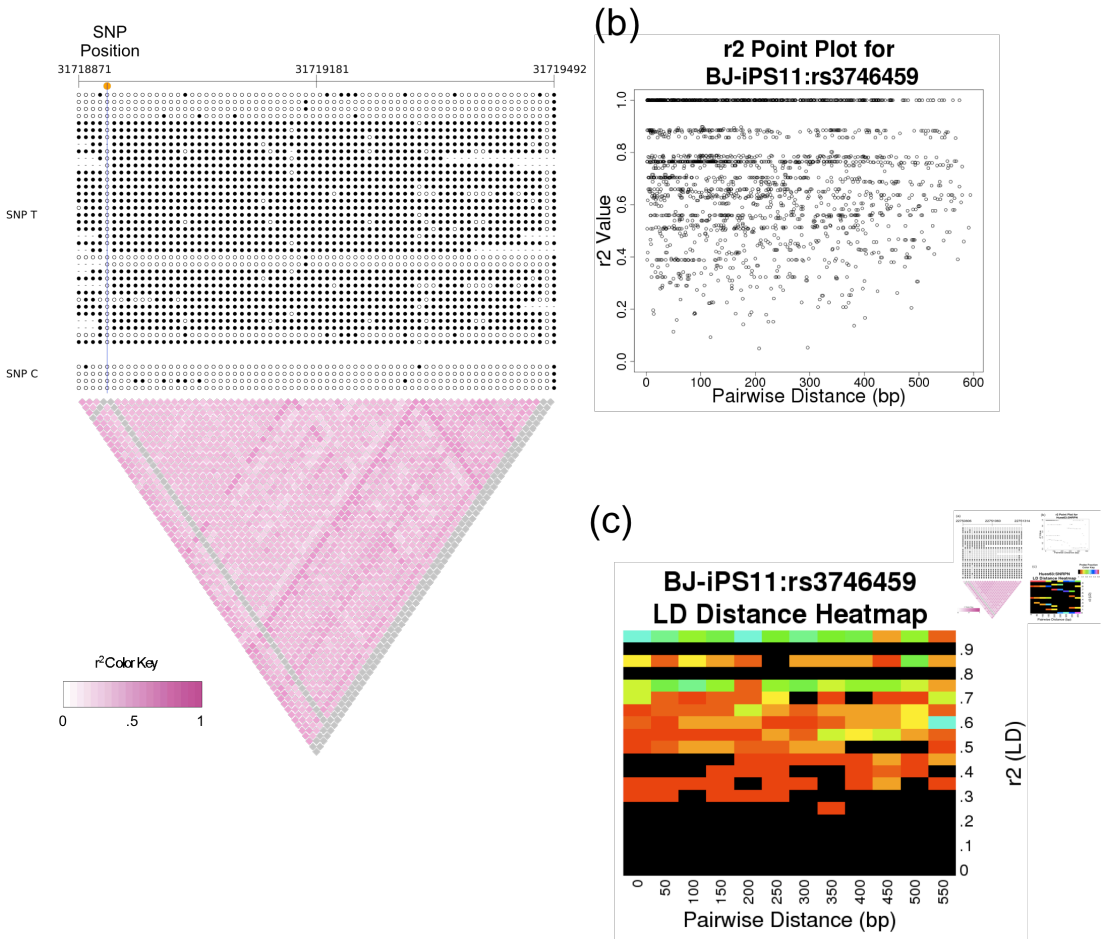


Figure 4.3.8 Linkage disequilibrium analysis of CpG methylation haplotypes around rs3746459 in BJ-iPS12 (chr20:31718871-31719492). (a) Sanger shows sequences that are either mostly methylated or unmethylated. In this case, the T SNP allele destroys the CpG dinucleotide and CpGs at this location on the T allele are thusly unmethylated. The methylation organization shows characteristics of ASM but the SNP identity at rs3746459 does not clearly separate the methylated and unmethylated sequences. (b) and (c) CpG pairwise r^2 value plots and heatmap show that the r^2 values in this region are significant and maintain significance as the relative distance between the two CpG dinucleotides increases.

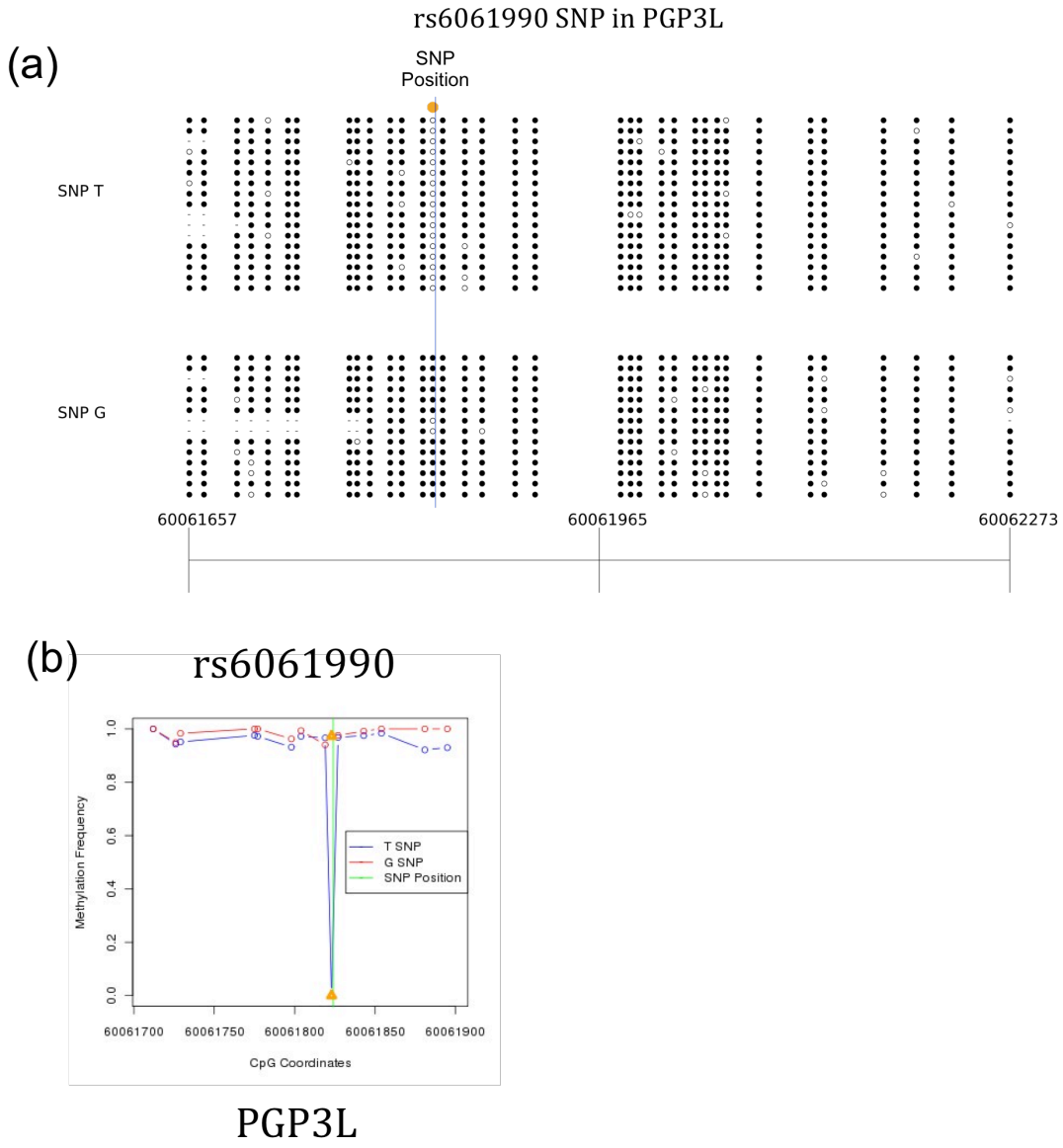


Figure 4.3.9 ASM example in PGP3L. (a) Sanger reads from PGP3L show ASM at a CpG dinucleotide that overlaps with SNP rs6061990 (chr20:60061657-60062273). The other CpG dinucleotides do not show ASM. (b) Methylation frequency diagram of the SNP rs6061990 region produced by Illumina data (chr20:60061712-60061895). The diagram supports the ASM behavior observed in the Sanger sequences. The orange triangles represent a SNP site that overlaps with a CpG. There is a difference in methylation frequency at coordinate 60061798 but it is not considered significant due to low read depth; only 6 mapped Illumina reads of the G allele cover this CpG dinucleotide.

Next we explored potential cis-regulatory mechanisms that account for ASM. We identified many ASM regions that contained at least one heterozygous SNP that overlapped with a CpG dinucleotide. As an example, the region containing the T/C SNP site rs10846023 in the BJ-iPS12 cell line shows ASM that spans over 500 bp. This SNP is present at a CG dinucleotide and the T allele disrupts the CG site. However, the ASM behavior at this site is also exhibited in nearby sites (Figure 4.3.7), which suggests that some uncharacterized regulatory mechanisms determine the ASM and such regulation may be dependent on the local sequence context in cis. We also found examples of specific methylation behavior that did not correlate with the base identity of a nearby SNP (4.3.8). Given the presence of SNP containing CpGs, we refined our ASM categories so that category I represented ASM that was not solely dependent on SNP containing CpGs. Category II represented ASM that was solely dependent on SNP containing CpGs. We found 45-77% of category I SNP regions have SNPs at CpG dinucleotides, which suggests that SNPs on CpG sites may have an impact on differential methylation establishment if the SNPs are located in regions where methylation regulators are involved. In contrast are the regions where the ASM affects only CpG locations overlapping with SNPs. One allele of such SNPs eliminates the CpG sites, thus prevents them from being methylated. One example is the region around SNP site rs6061990 in PGP3L (Figure 4.3.9) and PGP9L (Figure 4.3.7). The PGP9L cell line shows ASM at two CpG dinucleotides while the region in PGP3L shows ASM at one CpG site. This example demonstrates

that individual ASM sites can be dependent on sequence difference alone. A complete list of ASM categorizations of SNP regions is available on our supplementary website (http://genome-tech.ucsd.edu/public/ASM/). Overall 38~88% of ASM regions are solely due to the presence of SNPs at CpG dinucleotides, revealing that genetic variation at CpG sites is a dominating factor for ASM.

Finally, ASM regions that span across multiple CpG sites are likely regulated by other cis-regulatory mechanisms. We found many cases where the methylation state closely correlated with the alleles identified by a SNP. Such behavior is more likely explained by an allele-specific regulatory mechanism rather than cell subpopulations undergoing different epigenetic processes. For example, regions that exhibit reverse allelic preference in different cell lines of the same genetic background are observed (Figure 4.3.10), which can only be explained by a regulatory mechanism that involves more than one cis-regulator with opposite effects. Out of the category I ASM regions that are covered by our LD analysis, 8-35% are within methylation LD blocks, demonstrating that ASM analysis uniquely identified new regions where methylation was due to allele specific cis-regulation. Finally, the two IMR90 biological replicates showed that 83% of ASM calls were consistent in both experiments (201 matched ASM SNP calls out of 242 total ASM SNP calls). Therefore, ASM is due to biological regulation instead of biological noise or technical artifacts.

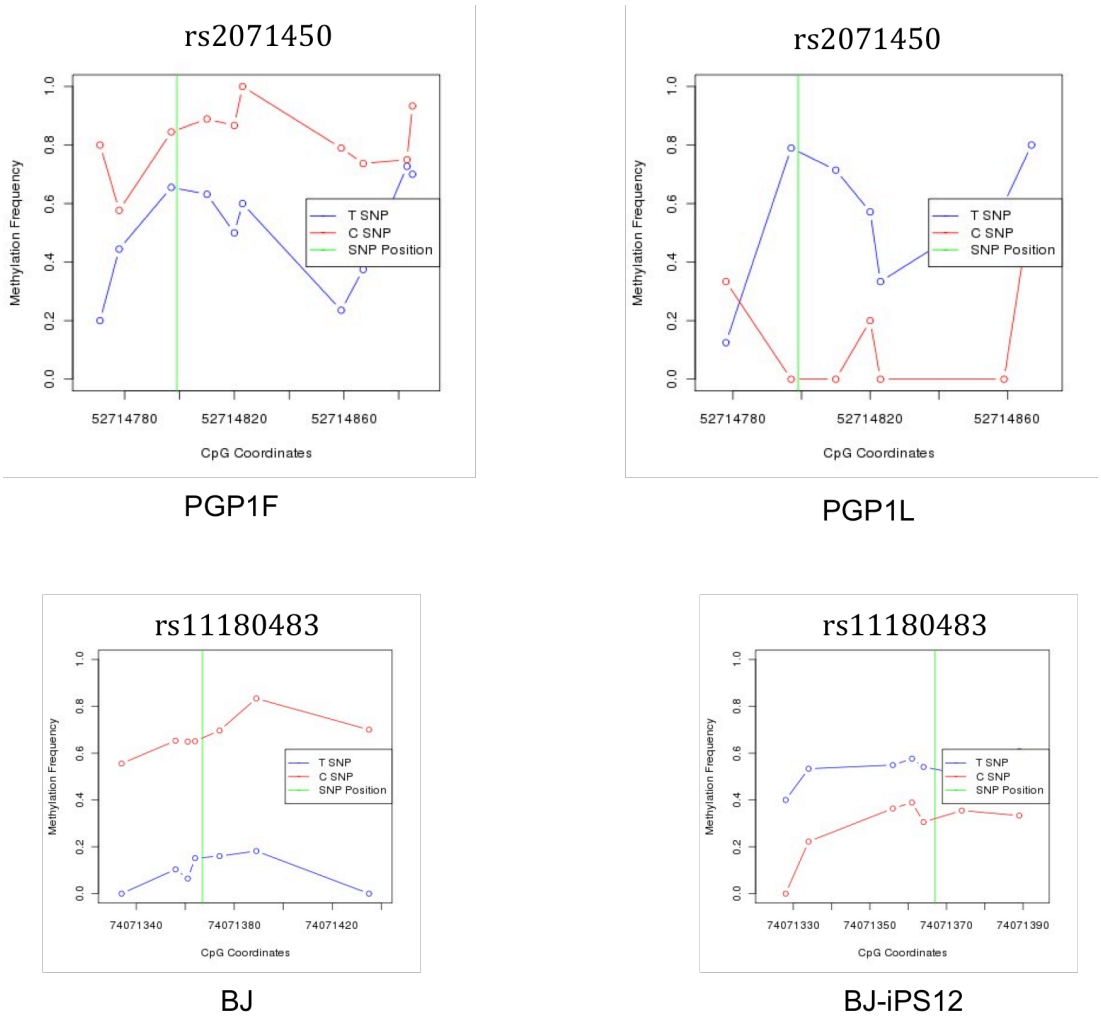


Figure 4.3.10 . Examples of ASM flipping. (a) and (b) Illumina reads from PGP1F and PGP1L show ASM in SNP rs2071450 region (chr12:52714771-52714885). The G allele is predominantly more methylated than A allele in PGP1F whereas A allele is more methylated than G allele in PGP1L. (c) and (d) Illumina reads from BJ (chr12:74071334-74071435) and BJ-iPS12 (chr12:74071328-74071389) show ASM around SNP rs11180483 indexed region. C allele is predominantly more methylated than T allele in BJ whereas T allele is more methylated than C allele in BJ-iPS12.

4.4 Discussion

We set out in this analysis to investigate the basis of fuzzy methylation with two novel approaches: adapting linkage disequilibrium analysis to methylation data, and performing SNP calling on bisulfite sequencing reads. This study represents the largest survey of ASM in the human genome to date. Hundreds of methylation LD blocks were identified from over 2,000 CpG islands in two human chromosomes and dozens of other regions. Roughly 30~48% of fuzzily methylated CpGs were found to have r^2 >0 .3 and 14~36% of fuzzily methylated CpGs were found within LD blocks. This shows stochastic effects do not explain a significant amount of observed fuzzy methylation. Our SNP ASM analysis found that 23 to 37% of heterozygous SNPs are associated with ASM. The frequency of ASM is similar to the frequency of allele-specific gene expression (ASE) observed in the human genome(*219-223*). ASE has been considered as an important indicator for the presence of functional cis-regulatory variants(*224*). ASE and ASM could be tightly coupled by the same cis-regulatory variants(*225, 226*). Similar to gene expression, methylation or ASM can be considered as quantitative traits for population genomic analysis. One important finding of this work is that a SNP could be a functional cis-regulatory variant by disrupting a CpG methylation site. According to the snp129 database, there are 225,659 known SNPs that locate on CpG sites. In addition, because CpG dinucleotides are highly mutable, there are many CpG rare variants in individual genomes(*227*). CpG

SNPs are likely an important class of cis-regulatory polymorphisms that connects genetic variation to the individual variability of the epigenome.

4.5 Methods

4.5.1 Bisulfite Targeted Resequencing with Padlock Probes

Bisulfite padlock capture and targeted resequencing were performed as described(*38*). Briefly, genomic DNA was extracted from frozen pellets of lymphocyte, fibroblast, iPS or hES cells using Qiagen DNeasy columns, and bisulfite converted with Zymo DNA methylation Gold kit (Zymo research). The Cpg30k padlock probe library was annealed to the bisulfite-converted sample DNAs, circularized, and amplified by PCR. Random shotgun sequencing library was generated by USER/MmeI enzymatic fragmentation from the amplicons of captured targets, and then were subsequently sequenced by Illumina Genome Analyzer.

4.5.2 Bisulfite PCR and Cloning for Sanger Sequencing

Bisulfite PCR reactions were performed in 100ul reactions including 50ng bisulfite converted genomic DNA, 200uM dNTP, 0.4 uM forward and reverse PCR primers, 1x IQ PCR Supermix (Bio-Rad) at 94º C for 2 min, 45 cycles of 94º C for 30 s, 62º C for 1min, 72º C for 1min, and finally 72º C for 5

min. Bisulfite PCR products were cloned in the PCR 2.1-TOPO vector (Invitrogen), and multiple clones were picked and sequenced at Agencourt. The primer sequences are listed in Figure 4.3.2.

4.5.3 Read Mapping

Single end 41 bp Illumina reads, which covered 11 cell lines, were merged with newer paired end 36 bp Illumina reads generated from 8 cell lines. Merged together, the combined data set represented 16 cell lines. The single end reads did not contain quality scores while the paired end data set did contain them (+64 offset). Single end reads were uniformly assigned 84 as a quality score for each base. We assigned this score so that the single end reads would pass quality filtering but were not given the same confidence as reads assigned higher quality scores by Illumina's software. For each cell line, the reads were sorted into two categories. The first category contained reads that represented bisulfite converted sequences and the second category included the reverse complementary reads. The ratio of T/C nucleotides was compared to the ratio of A/G nucleotides to determine to which category a read belonged. Reads that represented bisulfite converted sequences had a higher T/C ratio than A/G, since unmethylated cytosines were converted to thymines. Reads classified as reverse complementary were converted to bisulfite converted sequences (i.e. the reverse complement was taken). The reads were merged, demethylated in silico, and mapped to the unmethylated bisulfite converted hg18 genome. SOAP 2.20(*206*) mapped sequences to the

reference via an end-to-end policy that allowed up to two mismatches. Reads that mapped to multiple locations were discarded. For the paired end datasets, the paired ends were mapped independently of each other. Mates of paired end sequences that mapped to separate chromosomes were treated as single end reads. If both mates of a paired end sequence mapped uniquely, the mapped locations of these reads were checked. If the mates' sequences overlapped (i.e. distance less than 36 bp) or were greater than 250 bp from each other, the mates were treated as single end reads. For the remaining paired end reads, the distribution of distances between mates was calculated per cell line. The distributions were assumed to be normal and the distribution's parameters were calculated for each cell line. Paired end reads whose mate distances fell outside of 2.5 standard deviations from the average mate distance were treated as single end reads. The remaining reads were considered valid paired end reads. The generation of methylation frequency values for each targeted CpG was calculated analogously to the previously published protocol.

Sanger reads were mapped to a reference template using blat. Due to the much larger size of the sequence length and its known location on the genome, gaps and mismatches were allowed during the alignment. Analogous to Illumina reads, T/C and A/G ratios for Sanger sequences were calculated and if needed, Sanger sequences were reverse complemented so that they would match the reference template in the forward direction. The sequences

were then demethylated in silico and mapped to an unmethylated template using blat. The mapped read coordinates for each in silico demethylated read were then transferred to the original sequence. Sanger Sequence diagrams were created by aligning reads according to CpG sites and grouped together based on the nucleotide present at the SNP position.

4.5.4 SNP Calling

Heterozygous SNPs were detected using an algorithm that assigned each genotype a probability. For example, a diploid cell was tested for ten possible genotypes at a given nucleotide position: AA, AT, AG, AC, GG, GC, GT, CC, CT, and TT. The possible genotypes were bisulfite converted. The inputs for the SNP calling algorithm are the following: (1) reads that covered the nucleotide position, (2) the quality scores of those reads, and (3) the SNP129 dbSNP candidate positions and SNP identities. Base calls at the examined SNP site needed to have a minimum Phred-like quality score of 15 and the three flanking base calls on needed to have an minimum quality score of 15 on either side. Bisulfite conversion eliminates the complementarity between the Watson and Crick strands and the Watson and Crick strands were thusly treated independently. If reads mapped sufficiently to both strands, the bases at this position were examined to ensure that the bases on Watson and Crick strands were indeed reverse complementary. If a certain base was present in more than 20% of the Watson strand reads, its reverse complement needed to be present on at least 20% of the Crick strand reads. If

these criteria were not met, this location was not analyzed. For sites that passed this filter, a nucleotide frequency matrix was constructed for each nucleotide position based on the read data. The Phred-like scores were used to weight the nucleotide count contributions to the nucleotide frequency matrix. For example, a nucleotide call with a Phred-like score of 20 led to a 0.9 contribution while a nucleotide call of 30 led to a contribution of 0.999. This weighting scheme attenuated the effects from lower quality base calls. The weighted matrix was normalized so that the sum of all entries equaled 1. Each entry in the matrix was then multiplied by the read count.

A Fisher's Test was used to calculate the probability that the nucleotide frequency matrix represented a specific genotype. To test genotype AG, the Fisher's Test compared the frequency that A appeared in the read data to the expected frequency A would appear if this position were an A/G heterozygous SNP. A second Fisher's Test was performed that compared the frequency G from the read data to the expected frequency at an A/G SNP site. The two p-values were multiplied together to generate a stranded p-value product for a specific genotype. The product of the strand specific p-value products represented the likelihood of a specific genotype at a nucleotide position. The likelihoods of all genotypes were then normalized so that the sum of these likelihoods was 1. To filter out false positives, a SNP candidate site needed to have an odds ratio greater than 10 relative to the next most likely genotype. In the case of Hybrid1, the most likely genotype at a SNP candidate site needed to have an odds ratio greater than 100,000 relative to the second most likely

genotype. This stricter threshold was implemented due to SNP calling on a tetraploid cell line. SAMtools was not designed for tetraploid cells and we were unable to verify our SNP calls with SAMtools for Hybrid1. Only SNP sites reported by the SNP129 database were examined. Each examined SNP candidate site needed to have at least 10x read depth. With double stranded information, the SNP calling algorithm was able to discern the original SNP identity. Some single stranded SNP identities, however, could not be clearly identified with bisulfite converted reads. Since reads were demethylated in silico during the SNP analysis, C/T SNPs were not called and A/C and A/T SNPs were not resolved for single stranded SNP calls. However, since SNP calls were made at SNP129 sites, the single stranded SNP identity could be clearly discerned based on the allelic information in the SNP 129. SNP calls that were not consistent with the SNP129 database were excluded. If a called SNP created a CpG dinucleotide or destroyed a CpG dinucleotide relative to the reference genome, it was recorded. The methylation frequency of found CpG dinucleotides not present in the reference were investigated analogous to the method state above. These new CpG dinucleotide sites were used in the SNP ASM and LD analyses.

We also called SNPs using SAMtools v 0.1.7 in order to improve the confidence in SNP calls made with bisulfite converted read data. Mapped reads were demethylated and SAMtools created a consensus sequence. Examining sites with at least 10x read depth, we searched for base calls in the consensus sequence with a minimum SNP score of 20 that did not match the

reference. Analogous to our own SNP calling, SNP sites with double stranded coverage allowed for unambiguous SNP calling. We used the intersection of SNP calls between SAMtools to form a confident SNP candidate list, which we used in the other analyses.

4.5.5 ASM SNP Identification

A SNP site was labeled as an allele specifically methylated (ASM) region if the region met one of the following criteria: (1) there was a single CpG site that showed significant ASM, (2) all CpGs, when summed together, showed significant ASM, and (3) all non-SNP overlapping CpGs, when summed together, showed significant ASM. To calculate ASM, a contingency table was created where the columns represented alleles and the rows represented the cytosine (methylated) and thymine (unmethylated) counts at CpG sites found on those alleles. A Fisher's Exact Test was used to produce a two sided p-value, which served as the metric for ASM. For (1), each CpG site was treated independently and a contingency table was thusly made for each CpG site. For (2) and (3), one contingency table was made and the cytosine and thymine counts were summed across multiple CpGs per allele. If the p-value for any of these analyses was less then 0.001 and the methylation frequency difference between the alleles was greater than 0.1, the SNP region was labeled as ASM. The methylation frequency difference serves to filter out regions with very high read coverage (1,000+) that the Fisher Test would identify as significant ASM even though the allelic methylation frequency

difference would be quite small. There were three ASM categories. Category I ASM is not solely dependent on a SNP present in a CpG site while category II ASM is solely dependent on the presence of a SNP at a CpG site. Category III regions showed no ASM. For example, if a SNP region had a significant p-value in analysis (1) but the ASM CpG overlapped with a SNP, then this region was usually labeled as category II ASM. If a SNP site had a significant p-value in analysis (2) but not in (3) or (1), it was labeled as category II ASM. SNPs that had a significant p-value in (1) for CpGs that did not overlap with SNPs or had a significant p-value in (3) were labeled as category I ASM. Using the fdrtool in R, we calculated the FDR for our per CpG ASM and average ASM calls across all cell lines. A p-value cutoff of 0.001 yields a 0.62% FDR for our per CpG ASM calls and a 0.25% FDR for our average SNP ASM calls. Sanger sequences were labeled using the same criteria as the Illumina reads. Regarding the Illumina data, CpG sites with less than 5x read depth on either allele were not considered in the ASM SNP analysis. There was no minimum read depth requirement for the Sanger data.

4.5.6 LD analysis

The r^2 metric was adopted from linkage disequilibrium analysis to measure the organization of methylation at CpG sites on the same read sequence. Reads that contained more than one CpG were used in this analysis and the methylation status of all present CpGs was recorded. For the Illumina data, only CpG pairs that were covered by at least 10 reads were

considered in this analysis. For Sanger data, only CpG pairs covered by at least 3 reads were considered. For each recorded CpG pair, a contingency table was constructed, which counted the combinatorial methylation states of the CpG pair (i.e. both methylated, both unmethylated, or mixed methylation states):

Table 4.5.1 Table format used to calculate r^2 values.

CpG 1 / CpG 2	**Methylated**	**Unmethylated**
Methylated	$F_{Methyl\backslash Methyl}$	$F_{Methyl\backslash Unmethyl}$
Unmethylated	$F_{Unmethyl\backslash Methyl}$	$F_{Unmethyl\backslash Unmethyl}$

The r^2 values were calculated based on data in the contingency table via the following equation:

$$r^2 = \frac{\left(F_{Methyl\backslash Methyl}F_{Unmethyl\backslash Unmethyl} - F_{Methyl\backslash Unmethyl}F_{Unmethyl\backslash Methyl}\right)^2}{F_{Methyl\backslash *}F_{Unmethyl\backslash *}F_{*\backslash Unmethyl}F_{*\backslash Methyl}}$$

CpG sites that displayed uniform methylation patterns were not considered in this analysis since they did not produce meaningful r^2 values. LD blocks were created based on the presence of a CpG pair that contained a significant r^2 value (i.e. $r^2 > 0.3$). These blocks were extended if there was either an overlapping CpG pair with a significant r^2 value or there was a CpG pair with a significant r^2 value within 100 bp. Since we were interested in looking at regions of extended organized methylation, LD blocks less than 100 bp or containing less than 10 CpG pairs with $r^2 > 0.3$ were filtered out.

All raw data, SNP calls with ASM categorizations, LD analyses, Sanger diagrams, and methylation frequency data are available at http://genome-tech.ucsd.edu/public/ASM/.

Chapter 4, in part, is a reprint of the material as it appears in R. Shoemaker, J. Deng, W. Wang, K. Zhang, Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* **20**, 883 (Jul, 2010). The dissertation author was the data analyst and one of the principal authors of this paper.

Chapter 5 Implementation of the Methylation Analysis Pipeline

5.1 The Purpose of the Pipeline

The purpose of this methylation pipeline is to make the process of mapping and deriving biological information from Illumina BSC sequencing data more convenient. Currently there are no programs that offer a simple way to map Illumina BSC data to the genome, call SNPs on BSC sequences, and find regions that show interesting methylation patterns. The pipeline performs all of these tasks and only requires that the user upload his/her data and input a few parameters. Since the pipeline server performs all methylation processing, the pipeline frees up considerable computational resources for the user. The pipeline will ultimately output biologically relevant information, i.e. methylation frequency information, called SNPs, ASM regions, and regions showing organized methylation patterns. The outputted information enables the user to quickly perform inter- and intracellular analyses on his/her sequencing data.

Development of the pipeline started on smaller targeted BSC sequencing data sets but it is now designed to process whole genome BSC shotgun sequencing data. The pipeline achieves this capability by relying on MySQL to efficiently store large amounts of data to a hard disk where it can be quickly retrieved for analysis. The pipeline is thusly capable of handling billions

of reads that describe the methylation states of millions of CpGs in the human genome. A simple e-mail notification system notifies the user when her/his data is queued, processing, and when the output files are ready for download.

5.2 Pipeline Outline

The methylation pipeline performs 7 core functions: (1) mapping of BSC sequences to the human genome, (2) calculation of methylation frequencies for CpG sites found on the reference genome (3) SNP calling on the mapped reads, (4) discovery of created or destroyed CpG sites based on the presence of SNPs, (5) calculations of methylation frequencies for created CpGs, (6) categorization of SNPs into ASM classes, and (7) methylation LD analysis.

5.2.1 Mapping BSC Sequences to the Genome

Mapping algorithms such as Bowtie(*228*), SOAP(*206*), and BWA(*229*) greatly simplify the task of finding the chromosomal coordinates of next generation sequencing data. The trickiest aspects tend to be setting appropriate thresholds for the mappability of a sequence (e.g. how many times can a sequence map to the genome, how many base mismatches are allowed in a read, etc.). BSC sequences are more difficult to map because of the following two reasons: (1) possible cytosine methylation sites can be one of two bases (C or T after PCR amplification) and (2) BSC sequences are no longer reverse complementary.

To avoid mapping issues due to a cytosine's unknown methylation state, all reads are demethylated *in silico*. The reference genome is also converted so that the demethylated reads map properly. This entails creating two entries for each read sequence. The first entry contains the original sequence and the second entry contains the read sequence after all cytosines have been replaced with thymines. The second entry is used in the mapping the algorithm and the mapped chromosomal coordinates are then associated with the original sequence. The second entry reads represent *in silico* demethylated reads and these are of a uniform methylation state. The reference genome is also demethylated *in silico* so that demethylated reads can map to it. If these sequences were not demethylated *in silico*, sequences containing methylated cytosines would map to the demethylated reference genome with mismatches at the methylated cytosine locations. This would lead to a biased mapping of unmethylated reads. CpGs would thusly appear to be less methylated or their methylated states would be unknown due to the poor mappability of methylated reads. If the reads and the reference sequence were not demethylated *in silico*, reads would have sequences that contained thymines at unmethylated cytosines positions. In this case reads containing all methylated cytosines would map perfectly to the genome but reads with unmethylated cytosines would have mismatches at unmethylated cytosine positions. This would lead to a biased mapping of methylated reads. The simplest method to avoid biasing mappability towards unmethylated or

methylated cytosine states is to demethylate both the reads and the reference genome *in silico*.

There are two mapping strategies that take into account the loss of reverse complementarity after BSC. After BSC, unmethylated cytosines appear as uracil (or thymine after PCR amplification). Uracil and thymine do not base pair with guanine. To deal with this issue, the human methylome study created a customized reference genome(*1*). In this process, two separate reference genomes are created, one for the Watson strand and one for the Crick strand. The Watson strand genome is the reference sequence after it has been demethylated *in silico*. The reference Crick strand undergoes a different base substitution. All guanines on the reference Crick strand are replaced with adenines. Adenines replace guanines in the reference Crick strand because thymines replace cytosines in the reads as they are demethylated *in silico.* Reads that are on the Crick strand will map as their reverse complement to the reference AG-replaced reference strand (Figure 5.2.1).

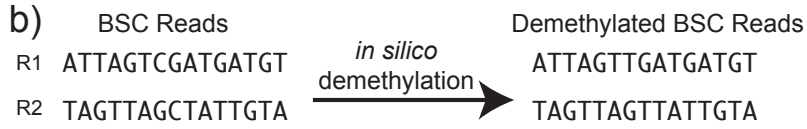
The second mapping strategy also creates two methylated genomes but it does not make any G to A substitutions. This mapping style is used in the BSPP studies(*38, 39*). Similar to the first method, two separate reference genomes are created, one for the Watson strand and one for the Crick strand. The Crick strand reference genome is the reverse complement of the reference genome. Both reference genomes are then demethylated *in silico*. The reads are then analyzed to determine whether they represent a BSC

sequence or the PCR-reverse complementary product of a BSC sequence. This determination is made based on the ratio of C/T and G/A bases. A BSC sequence is expected to contain a lower ratio of cytosines to thymines than guanines to adenines. A PCR product that uses a BSC sequence as a template is expected to have a higher C/T ratio than G/A ratio since the PCR product is reverse complementary to the BSC sequence. If a read is determined to be a BSC sequence, it is demethylated *in silico* and mapped to both reference genome strands. If a read is determined to be a PCR product, it is reverse complemented, demethylated *in silico*, and then mapped to both reference genomes (Figure 5.2.1).

My pipeline utilizes both techniques since each technique offers different advantages. The first technique does not involve taking the reverse complement of read sequences. Most next generation sequencing data tends to have an error bias towards the 3' prime end of a read sequence since data can become more unreliable as the sequencing cycle count increases. Mapping via the first method allows the mapping algorithm to set stricter sequence alignment thresholds for the 5' end than the 3' end. This also allows post-mapping sequence trimming of the 3' end if it contains too many mismatches. However, this method does not optimally map PCR products from a BSC template. The second method maps both BSC and PCR produced sequences but many of the reads are reverse complemented before they are mapped to the reference strands. If a next generation read data set is noisy,

the first method is more appropriate while the second method is better for a data set with a consistent quality score across most of the read.

Figure 5.2.1 a) Double stranded DNA loses reverse complementarity at unmethylated cytosine positions during bisulfite conversion. The lines between the strands represent base pairing. In this figure, Crick strand sequences are shown 3' to 5' to show complementarity with sequences on the Watson strand. b) Reads are demethylated *in silico* before mapping. This process involves replacing all cytosines with thymines. R1 and R2 represent two independent sequencing reads. c) The GA mapping method uses two reference genomes: one reference genome is *in silico* demethylated while the other reference genome undergoes a gunaine to adenine conversion. BSC reads from the Watson strand will map to the *in silico* demethylated reference genome while reads from the Crick strand will map to the GA converted genome after reverse complementation. d) As an alternative to the GA method, the strand detection method detects whether a read represents the BSC read or the PCR product of a BSC sequence. PCR occurs in the absence of bisulfite and the presence of cytosines on the PCR products thusly do not necessarily represent methylated sites. Reads that have a higher A/G ratio than T/C ratio are labeled as PCR products of BSC sequences. These sequences are reverse complemented before *in silico* demethylation occurs (not shown). Then the *in silico* demethylated reads are mapped to the reference genome. There are two reference genomes used in this method. The first one is an *in silico* demethylated Watson strand and the other is an *in silico* demethylated Crick strand. All reads map to either reference genome in the forward direction (i.e. all reverse complementation occurs before mapping).



c) GA Method

Reference Mapping Sequences

Watson ATTAGTTGATGATGT

Crick ATCAATCAATAACAT

Demethylated BSC Reads

R1 ATTAGTTGATGATGT

R2 TAGTTAGTTATTGTA

Reverse Complement of R2 ATCAATCAATAACAT

R1 Alignment

Watson ATTAGTTGATGATGT

R1 ATTAGTTGATGATGT

R2 Alignment

Crick ATCAATCAATAACAT

Reverse Complement of R2 ATCAATCAATAACAT

d) Strand Detection Method

BSC Reads

R1 ATTAGTCGATGATGT

R3 ATCAATCGATAACAT

	BSC Reads	A/G Ratio	T/C Ratio	Result
R1	ATTAGTCGATGATGT	1	6	Demethylation
R3	ATCAATCGATAACAT	7	4/3	Reverse Complementation Before Demethylation

Reference Mapping Sequences

Watson ATTAGTTGATGATGT

Crick TAGTTAGTTATTGTA

R1 Alignment

Watson ATTAGTTGATGATGT

R1 ATTAGTTGATGATGT

R2 Alignment

Crick TAGTTAGTTATTGTA

R3 Reverse Complement TAGTTAGTTATTGTA

This pipeline also takes into account clonal reads. Clonal reads are considered in two ways: (1) reads whose sequences are identical and (2) reads that map to the same chromosomal location. Given that sequencing error can produce two slightly different sequences of reads based on an identical template sequence, the latter category is used to commonly define clonal reads. However, the presence of SNPs can lead to a false identification of clonality when using the second definition. In experiments where fragmentation of the BSC human genome is considered random, the likelihood of a finding two sequences that share the same starting location is much less than the likelihood of two PCR products produced from the same sequence. Since PCR amplifies sequences non-uniformly, the distribution of sequences after PCR amplification can be different than the original distribution. Clonal filtering alleviates any introduced bias from PCR since over-amplified PCR copies are filtered out of the data set. Since the pipeline can process BSC shotgun sequencing data, which can contain several billion reads, these sequencing reads may come from separate sequencing libraries. Each sequencing library represents a separate library construction experiment. Reads that share the same start coordinate within a sequencing library are considered clonal while reads with the same start coordinate in different libraries are not considered clonal. Association of the sequencing data with sequencing libraries is established by the experimentalist who uploads a file containing this information with his/her next generation sequencing data. There are experiments, however, where fragmentation of DNA is not random,

which decreases the likelihood that two original sequences sharing the same start coordinates are clonal sequences. The user can thusly decide whether to enable clonal filtering depending on his/her experiment.

The mapping part of the pipeline ensures the read maps uniquely. It also checks if the distances between the paired sequences are within a specific range. Since both mapping strategies involve two reference genomes, the mapping algorithm has to be executed twice. All of the mentioned algorithms include parameters that can filter mapped reads so that only uniquely mapped reads are reported. A uniquely mapped read is a read that has a best match (i.e. least amount of mismatches) at a single location on the genome. However, since the mapping algorithms are executed twice, some reads may map uniquely to both copies of the customized reference genomes. The pipeline performs additional filtering after read mapping to ensure that all reported reads mapped only once to either version of the customized reference genomes.

The pipeline can also process paired end data but additional user-defined parameters are required. Paired end sequences improve ASM- and r^2-based analyses since the reads cover 2x more of a single molecule than single end sequencing. Mapping algorithms have built-in capabilities for paired end analyses and these analyses do not apply well to BSC sequencing data. In paired end sequence samples where one sequenced end is reverse complementary to the other, the sequence ends will map to separate customized reference sequences. In this case, the mapping algorithm maps

each end independently from the other end. Only after read mapping has completed are the two ends verified as an appropriate paired end sequence. With non-BSC data, a major mapping advantage of paired end sequences is the ability of a paired end to map uniquely even if one or both sequencing end(s) do not map uniquely to the genome; the mapping algorithm checks if the combination of both ends maps uniquely to a certain region. However, since the ends are mapped separately with BSC data, each end must map individually to the genome. This leads to the exclusion of some paired end reads. However, as read lengths approach 100bp for Illumina sequencing, which is the most common sequencing technology, the issue of reads mapping to multiple locations is diminished. Additional issues within the sequencing library generation steps may lead to incorrect paired end sequences. Paired end artifact sequences can be filtered out by requiring a certain bp distance between sequencing ends. This distance can be determined based on the expected length of the investigated DNA fragments. The user specifies these minimum and maximum bp distances. The user does not need to specify whether his/her dataset is paired end; the pipeline automatically detects paired sequences.

5.2.2 Differences in Mapping Algorithms Used in the Pipeline

The pipeline allows the user to use either Bowtie or SOAP for mapping sequences. Bowtie is much more customizable while SOAP maps sequences more reliably. As mentioned in the previous section, paired end sequence ends are mapped independently of each other. Some ends may not map uniquely but the combination of both ends may lead to a unique match. Bowtie allows the reporting of multiple matches up to a certain threshold set by the user. The pipeline allows bowtie to report up to 100 matched locations for each end. After both ends have been matched, the pipeline then searches for a unique region where both ends are within an appropriate distance from each other. If such a region is found, the sequences are considered paired and their mapped coordinates are recorded. Unfortunately, SOAP does not allow the user to specify a matched location threshold; either SOAP reports only unique hits, all hits, or it chooses a random hit. A hit is an acceptable chromosomal coordinate for a read. Since some reads map thousands and thousands of times to the genome, it is not feasible to allow SOAP to report all hits when processing billions of reads.

In addition to precisely controlling the number of reported matches for a read, Bowtie allows for the user to specify the number and type of mismatches allowed in the sequence. SOAP allows the user to align sequences allowing

up to a certain number of total mismatches. Bowtie allows the user to align allowing: (1) a certain number of total mismatches or (2) a seed of the sequence, usually 28bp from the 5' end, up to a certain amount of mismatches. Unlike SOAP, Bowtie supports the Maq read mapping policy. The Maq policy entails mapping a seed with at most a certain number of mismatches and then allowing mismatches in the tail sequence until the sum of the mismatched bases' quality scores reaches a certain threshold. The Maq policy allows more mismatches with low quality scores than mismatches with high quality scores.

Although SOAP's read mapping policies are not as flexible as Bowtie's, SOAP maps slightly more sequences than Bowtie (Table 5.2.1 and Figure 5.2.2) using the same mapping policy. Bowtie 0.12.5 appears to discard many reads based on its belief that the sequence does not uniquely map to a single region (i.e. too many matches). SOAP, however, finds that many of these same sequences do uniquely map back to the genome. BLAT, another sequence aligning program, verified SOAP's results as correct since it found only one best match for the sequences that SOAP mapped but Bowtie did not(*230*). Bowtie offers the best parameter control for mapping but SOAP maps more BSC sequences. Given that both of these algorithms offer something unique, they are included as options. The experimentalist decides which mapping algorithm is most appropriate for his/her sequencing data.

Table 5.2.1 This table shows the mapping output from Bowtie and Soap using the same input data set. Both algorithms were set to report reads that mapped only once to the genome. Bowtie was executed using the strata, tryhard, and best options. Both sequences were mapped allowing at most 2 mismatches total (-v 2 option for both algorithms).

Mapping Algorithms	Reads Mapped	Total Reads	Mapped Percentage
SOAP 2.0	2,849,975	14,795,567	19.26%
Bowtie 0.12.5	2,663,293	14,795,567	18.00%

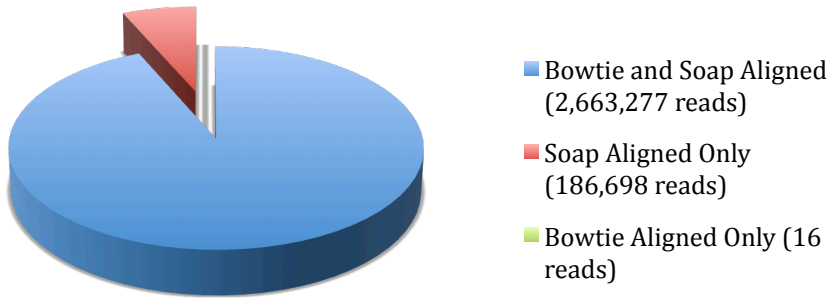


Figure 5.2.2 Soap maps around 1% more reads than Bowtie, which amounts to 186,698 reads in the sample dataset, which are 36 bp BSC Illumina read sequences. Bowtie maps 16 reads uniquely while SOAP does not. Assuming these percentages scale linearly, SOAP will map an additional 12.6 million reads for every billion input reads. The mapping algorithm parameters are explained in Table 2.5.1.

5.2.3 Mapping Alternatives

Given the presence of sequencing technologies other than Illumina (e.g. 454 and SOLiD) and the diversity of experiment types, the pipeline allows users to upload sequences that they have mapped. This is an important feature for the methylome data produced by Joe Ecker's group, for example. Since that group was able to produce 87bp long reads when most other labs were producing 36bp long reads, the sequencing quality towards the 3' end was quite low. To cope with this issue, the group performed trimming before and after mapping to the genome to eliminate sequence segments that were suspected of containing a significant amount of noise. Other labs may also experience experiment-specific sequencing issues and have thusly developed a customized mapping policy to account for them. The pipeline is still useful for these labs since they can upload their mapped data and find organized methylation regions including ASM, SNP calls, and methylation frequency information. This is a welcome alternative for groups who do not want to be confined to the methylation pipeline's mapping options.

5.2.4 Calculating CpG Frequencies

With the original read sequences and their associated chromosomal coordinates stored in a MySQL database, the frequencies are calculated by looking at each CpG site and counting the occurrences of cytosine and thymine at that position. There are over 28 million CpG sites in the human

genome and interrogating the methylation status of each of them is a computationally intensive process. To expedite this process, I have created an algorithm that breaks each chromosome into 250 million bp chunks. For each chunk, the reads that are located within each chunk are stored into memory. Once loaded into memory, the sequences are scanned for CpG coverage. If a read covers a CpG site, the base identity at that position is stored into another database. Performing these operations for chunks of chromosomes allows for the read information to be stored into memory and also reduces the number of MySQL queries. If all reads that mapped to a certain chromosome were stored into memory for a BSC shotgun sequencing experiment (e.g. 50 to 200 million reads per chromosome), the required memory would be greater than 32GB. If reads were not read into memory before CpG searching, then billions of MySQL queries would be made as each read is examined and this would be a slow process.

Once the base identities at chromosomal positions are known, the final step is to compute the fraction of bases that are cytosines (methylated) and thymines (unmethylated). The pipeline first indexes all CpG coordinates that are covered by at least one read. Each CpG coordinate is scanned and the base identities are stored into a hash. In order for a CpG site to be considered reliably mapped, at least 90% of bases at this position must be cytosine and thymine. Then the number of cytosines found is then divided by the sum of cytosines and thymines found at each CpG site. This final value is considered a methylation frequency and it is accompanied by a read count. Read counts

bestow a confidence level onto the methylation frequency. For lower read counts, an erroneous base call will more greatly affect the methylation frequency value. To limit this possible source of noise, methylation frequencies with a read count of less than 5 are not reported.

5.2.5 SNP Calling

The SNP calling algorithm generates genotype probabilities (diploid genotypes are: AA, AT, AG, AC, TT, TC, TG, CC, CG, and GG) and assigns a SNP candidate site the most likely genotype. This custom algorithm is detailed in Section 4.5.4 but additional changes were made as it was incorporated into this pipeline. The biggest change was the expansion of the SNP calling algorithm so that it makes calls at all dbSNP 130 sites. SNP Sites that are listed in the dbSNP 130 database are considered possible candidates. In the published version of the SNP calling algorithm, quality scores were taken into account when making SNP calls. Based on that Cpg30k dataset, I found that quality scores did not act as a filter for false positive SNP calls; the quality score component did not help the results and it slowed down the algorithm. I thusly skipped quality score considerations in order to accelerate the SNP calling algorithm. The minimum read depth threshold is 10x. To further optimize the SNP calling script, I cached Fisher's Test p-values so that Fisher's Test calculations would not have to be recalculated if the same Fisher's Test was recently performed. I used this SNP calling algorithm to call SNPs on 26 different samples that were targeted via the Cpg97k probe set.

Independent array-based SNP calls had also been made on these samples. I had 98-99% SNP call concordance rates, which were similar to the previously published SNP concordance rates for the Cpg30k data in Chapter 4.

To eliminate false T/C SNP calls due to the presence of methylation and unmethylated cytosines at a single site, all reads are first demethylated *in silico* before they are analyzed for SNPs. Due to this demethylation step, T/C SNP calls cannot be made on the T/C strand data since they cannot be distinguished from C/C or T/T genotypes. However, T/C SNP calls can be made on double stranded data or if the reads map to the opposite strand of the T/C SNP since the reverse complement of T/C is A/G. Double stranded SNP calls are typically more accurate than single stranded SNP calls because the SNP site is validated on two strands. However, double stranded A/G SNP calls do not benefit from double stranded coverage since the other strand consists of a single base, T. A/G or T/C double stranded SNP calls are thusly considered as accurate as single stranded SNP calls.

5.2.6 SNP-Affected CpG Sites

Many SNP sites overlap with CpG sites and the pipeline thusly searches the CpG sites that are created or destroyed by these called SNPs. SNPs can affect a CpG in four ways: (1) a SNP can create a CpG at position 0 of a CpG site (the cytosine position on the Watson strand), (2) a SNP can destroy a CpG at position 0, (3) a SNP can create a CpG at position 1, and (4) a SNP can destroy at CpG at position 1 (Figure 5.2.3). In order to examine

these four possible effects, the reference sequence around each SNP site is stored into memory ([-1,+1] bp around SNP site). In order for a SNP site to destroy a CpG site, the CpG site must exist in the reference sequence. Thus the retrieved SNP sequence is checked for the presence of a CpG at position 0 or position 1. If a CpG exists at position 1, the SNP affects the guanine base of the CpG. This means that the cytosine on the Watson strand will always be unmethylated when the guanine base substituted for another base. When the guanine is substituted for another base on the Watson strand, the cytosine position of the CpG on the Crick strand is occupied by another base. A SNP at CpG position 1 affects the Watson and Crick CpGs analogously.

The zygosity and position of a CpG-destroying SNP affects the methylation frequency calculations differently. A homozygous SNP at CpG position 0 SNP will destroy the CpG on both alleles with a base substitution at the cytosine on the Watson strand and the guanine position on the Crick strand. If this CpG is covered only by reads that map to the Crick strand, the CpG will appear to be completely unmethylated. Since the CpG site is destroyed by the SNP, this would not be an accurate methylation measure. The presence of an unmethylated cytosine on the Crick strand reflects a change in genomic sequence and is not a product of CpG methylation regulation since the CpG has been destroyed by a SNP. Alternatively, if a homozygous guanine SNP follows a cytosine base, the SNP creates a CpG site.

Figure 5.2.3 a) The presence of a SNP at the cytosine position and guanine position of a CpG site on the Waston strand. The base used for methylation frequency calculations is the base at CpG position 0 on the Waston strand and the base at CpG position 1 on the Crick strand. Sequences in a) are not bisulfite converted. b) Given a homozygous SNP at CpG position 0, a methylation frequency is not computed on the Waston strand (too many A or G bases) and the methylation frequency is 0 on the Crick strand. c) Given a heterozygous SNP, the methylation frequency for the Crick strand is .5 and the methylation frequency for the Watson strand is not calculated. In this example, the intact CpG site is completely methylated but the presence of the non-reference base on one allele reduces the Crick strand methylation frequency. Methylation frequency calculations are averages across both alleles.

A heterozygous SNP that overlaps with a CpG site will affect only one of the alleles and thusly creates a CpG site or destroys a CpG site allele specifically. For example, a SNP that occurs as position 1 of a CpG site, which is labeled as guanine in a reference, will destroy the CpG site on the allele where the SNP base does not match the reference. Methylation frequency calculations performed on reads from the Watson strand will show a completely unmethylated CpG. Methylation frequency calculations on the Crick strand may reveal a significant amount of bases that are neither thymine nor cytosine and a methylation frequency calculation would be skipped. An interesting case is the presence of a T/C SNP at position 0 of a CpG site regardless of zygosity. If this were a heterozygous SNP and only the Watson strand were targeted, the reads from the T allele would appear to be to completely unmethylated and reads from the C allele could be methylated or unmethylated. In this case, the methylation frequency would be artificially lower than expected since about half of the reads would show the site to be unmethylated. Given these effects, SNPs that affect CpG sites as the sequence level are discovered and stored into a database.

5.2.7 Allele Specific Methylation Analysis

This analysis classifies SNP sites based on the methylation patterns of nearby CpGs. There are four classes of ASM in this analysis that are intended to separate SNP-CpG relationships into three categories: (1) SNPs that exhibit non-sequenced based ASM, (2) SNPs that exhibit sequenced-based ASM,

and (3) SNPs that exhibit no ASM behavior. The first category is subdivided into two subcategories so that a set of high confidence ASM sites can be produced. The purpose of this labeling scheme is to identify biologically regulated ASM and sequenced based ASM SNP regions.

Significant SNP-ASM relationships are determined in two ways: (1) individual SNP-CpG methylation state relationships, (2) the relationship of a SNP to the methylation states of multiple CpG sites. The first method directly looks at the CpG's methylation state on a read sequence that also covers the SNP site of interest. If a CpG is significantly more methylated for one SNP type than the other, the ASM relationship is considered significant. With large read depths, the first method would be the only method necessary to find ASM relationships. However, some experiments have lower coverage, which makes it more difficult to statistically assess ASM. To cope with the issue of read depth, the methylation states of all CpGs on a read that also covers the SNP site of interest are summed. For example, if both alleles for a certain SNP site had only 3x coverage, each CpG SNP calculation will have only three samples for each allele. If three CpG sites appear on the same read as the SNP site of interest with 3x coverage, summing across all three CpGs increases the sample size to 9. If these CpG sites were oppositely methylated in an allele dependent manner, then the ASM p-value (based on a Fisher's Test as described in Chapter 4.5.5) would be $< .1$ for a sample size of 3 and $< .001$ for a sample size of 9. In order for method 2 to find ASM, the sum of methylation states across all CpGs within the SNP region must show significant ASM

behavior. If the read depth is low and only one CpG site exhibits ASM, my algorithm will not find this instance of ASM behavior to be statistically significant.

The ASM category that includes SNPs exhibiting non-sequence based ASM has the highest potential to contain regions of allele-specific regulatory activity. For each SNP site in this group, the methylation state of at least one CpG or the summation of methylation states across multiple CpGs shows significant ASM. To eliminate the possibility of sequence based ASM (i.e. SNP overlapping CpGs), these CpGs may not overlap with any called SNPs. To create a subgroup of high confidence ASM candidates, SNPs exhibiting non-sequence ASM across CpGs that do not overlap with any dbSNP 130 sites are labeled as class I. Other CpGs that show SNP-based ASM but whose CpGs overlap with dbSNP130 sites are classified as class II. Class II ASM may still be caused by sequence based ASM since there could be an uncalled SNP present in a CpG. A common reason for a missed SNP call is insufficient read depth; my SNP calling script requires a read depth of 10x. Since the ASM behavior seen in this category is unlikely due to sequence variation, it is more likely due to a biological mechanism (e.g. imprinting). The regions in this category are the most appropriate for identifying allele specific markers and studying the mechanisms of ASM.

Class III contains SNP regions that exhibit sequenced based ASM behavior. All regions in this category contain a called SNP that overlaps with a CpG site. However, the SNP site of interest is not necessarily the SNP site

that overlaps with a CpG. Another SNP that is on the same haploblock as the SNP of interest may overlap with a CpG. For example, there are two SNP sites where SNP site 1 is G/C and SNP site is 2 A/T. Whenever one allele contains G at SNP site 1, SNP site 2 on the same allele always contains A. If SNP site 2 overlaps with a CpG, SNP site 1 could be categorized as class III due to the second SNP site. SNPs that create or destroy CpG sites affect the ability of a cytosine to become methylated. In this case, the methylation state is not dependent on biological regulation but solely on the adjacent genomic sequence. In order for a CpG overlapping SNP site to show ASM, the cytosine must be methylated at some frequency when the CpG site is intact. Consider a case where the second position of a CpG is affected by a heterozygous SNP. If the cytosine at this affected CpG were entirely unmethylated on the intact CpG allele, its methylation frequency would be the same as the CpG located on the second allele (Figure 5.2.4). Given that most of the human genome is methylated, destruction of CpG sites due to SNPs will be observed as ASM.

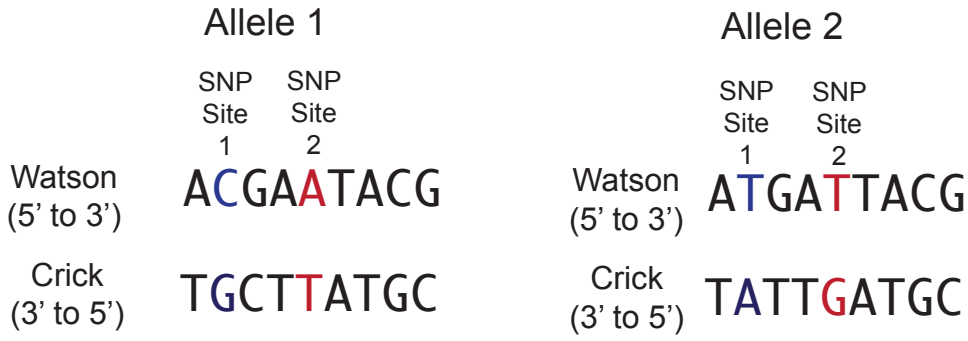


Figure 5.2.4 BSC sequences from a region that contains two nearby heterozygous SNPs. SNP site 2 serves as the reference SNP site for ASM analysis. Both SNPs are in phase (i.e. if SNP Site 1 is cytosine, then SNP site 2 is always thymine). Allele 1 contains the reference bases for both SNP sites. There are two CpG sites located in the region and SNP site 1 overlaps with the first CpG site. Looking at the Watson strand only, SNP Site 1 and CpG 1 exhibit ASM behavior. Since SNP 1 is a C/T SNP, it is impossible to discriminate between a SNP and biolgocial methylation variability given Watson strand sequences. Double stranded sequencing information distinguishes between the two possibilities.

SNPs for which there are no CpGs on the same read are considered class IV. SNPs that exhibit ASM behavior but the read depth is insufficient for statistical significance are also considered class IV. Class IV SNPs should not be considered true negatives since this class can contain sites that would clearly show ASM if additional sequencing data were available. Even if there are no CpGs on the same read as the SNP, CpGs sites might be located nearby (e.g.100-500 bp away) that show ASM. The purpose of this labeling scheme is to identify true positive ASM SNP regions. Additional work is required to identify true negative ASM SNP regions from the remaining SNP regions.

5.2.8 Methylation r^2 Analysis

The purpose of methylation r^2 analysis is to identify organized methylated regions, which include ASM regions, in cases where SNPs are not present to separate the alleles. The method is described in more detail in Section 4.5.6. The general outline of this analysis is to find the pair wise correlation of methylated states for CpGs that are located on the same read. Though reads are only 70-100 bp in length, overlapping reads can be used to daisy chain CpG sites to from larger LD block regions (Figure 5.2.5). LD blocks contain significantly correlated methylation states of CpGs and these LD blocks can be linked to biological regions. For example, the pipeline ranks RefSeq genes based on the presence of LD blocks contained by the gene; genes with the most LD block coverage are likely candidates for ASM activity. The r^2 value reflects how correlated these methylation states are, with $r^2 = 1$ representing strong methylation state correlation and $r^2 = 0$ meaning no methylation state correlation. Graph 5.2.1 shows how the r^2 metric corresponds to changes in methylation state correlation between two CpGs.

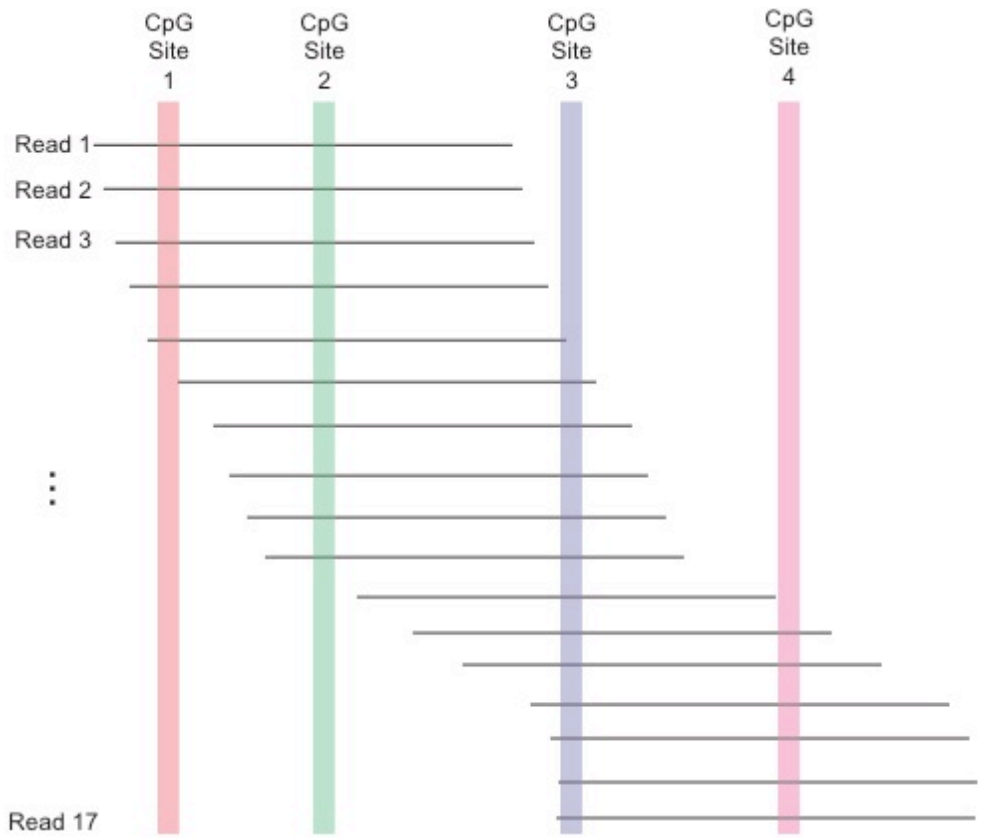
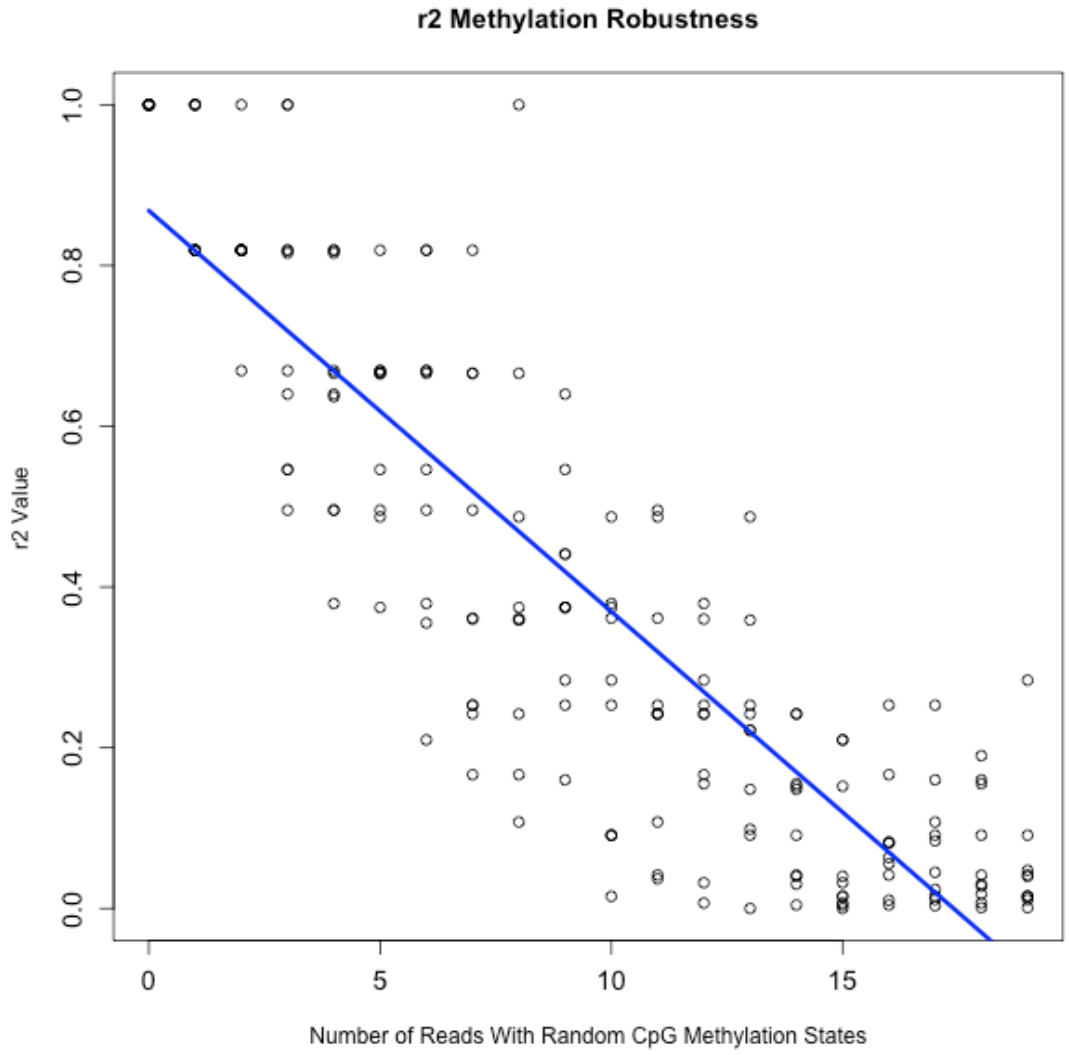


Figure 5.2.5 CpGs across different reads can form a single LD block. Each horizontal line represents a read and each vertical column represents a unique CpG site. CpG sites 1 and 2 are found on the same reads. CpG sites 2 and 3 are found on the same reads and CpG sites 3 and 4 are found on the same reads. If CpG sites 1 and 2 have a significant pairwise methylation correlation and CpG sites 2 and 3 have a significant pairwise methylation correlation, then CpG sites 1, 2, and 3 can be considered apart of a single candidate LD block. CpG site 4 can be included in the LD block if it significantly correlates with the methylation state of CpG site 3.

Graph 5.2.1 This graph shows how the introduction of random methylated states affects the r^2 value. The blue line is the linear regression trend line. This analysis involves a single CpG pair covered by 20 reads. Without the introduction of random methylation states, the methylation states of the paired CpGs correlate perfectly, which is represented at x=0. When randomly chosen methylated states for both CpGs on a certain number of reads are introduced into the dataset (x-axis), the r^2 value decreases. The x-axis represents how many of the 20 reads have been substituted for random methylation states of both CpGs. When x=20, all reads consist of random methylation information. Substitution of perfectly correlated methylation states for random ones was repeated 10 times and the data points for all 10 instances are plotted below; each x value has 10 y values. In this case, a CpG pair needs to have at least 8 reads that contain perfect correlation between the CpG's methylation states to have an r^2 value greater than .3. .3 is a threshold used for determining whether a CpG forms an LD block.



The pipeline produces r^2 values across all CpGs that appear on the same sequencing read with sufficient read depth. These values are then used to produce LD blocks and LD block gene coverage information. The LD blocks analysis is described in Section 4 with a single modification: CpGs had to have a methylation frequency between .25 and .75 (i.e. fuzzily methylated). This restricts LD blocks to regions where CpGs are found in both methylated and unmethylated states; the LD block analysis then describes the level of methylation organization at these fuzzily methylated regions.

The LD block gene coverage information is new to the pipeline. Once the LD block analysis has completed, all regions with sufficient read depth for LD analysis are mapped to RefSeq gene regions. A LD gene coverage value is calculated by firstly multiplying the number of significant r^2 values within LD blocks by the number of CpGs within the LD blocks. This product is then divided by the number of r^2 values not within LD blocks times the number of CpGs not located within the LD blocks (Figure 5.2.7). This value represents how many of the paired CpGs exhibit significant methylation organization. The values are weighted by CpG counts since one CpG can be involved in many pairwise r^2 calculations. If an LD block contains more CpGs, it is weighted more heavily. These LD block gene coverage values are then ranked. Genes that appear with the highest rank are considered promising candidates for biological activity that is associated with DNA methylation.

$$\frac{Significant\ r^2\ values\ in\ LD\ blocks * CpGs\ within\ LD\ blocks}{r^2\ values\ outside\ of\ \ LDblocks * CpGs\ outside\ LD\ blocks}$$

Figure 5.2.7 The equation used to compute LD block gene coverage values. The specified CpGs are within a RefSeq gene's annotated region ([+2kb TSS, -2kb 3' end]).The denominator involves CpGs outside of LD blocks but with sufficient coverage for the LD analysis.

I adjusted the methylation r^2 algorithm that I originally used in the Chapter 4 so that it would efficiently run on BSC shotgun sequencing experiments. For the targeted probe sets, all pair wise CpGs were scanned and their methylation states on reads the covered both sites were stored into memory. Similar to the CpG frequency calculations, storing all this information into memory is not feasible when analyzing billions of reads. The LD algorithm now scans CpG regions covered by mapped reads starting at the first covered CpG site per chromosome. The pair wise CpG methylation information is stored into memory until the algorithm encounters a read coverage gap or a CpG gap of 500 bp or more. When these gaps occur, the extension of an LD block is not possible. These gaps act as stop points where the algorithm processes the data it has stored into memory, outputs its results into a database, and erases the stored data. The algorithm then continues to the next CpG site. Unlike in the CpG frequency calculation program, the chromosome cannot be easily broken into equally sized chunks; a CpG pair could exist such that one CpG exists within the chunk and the other is located just outside the chunk's region. LD analysis must split the chromosomes into regions that are separated by large gaps so that same read CpGs are not

broken up into different chunks. This algorithm breaks up the chromosome chunks that preserve all pair wise single molecule CpG methylation information and thusly consumes a much smaller amount of memory.

5.3 User Interface For the Methylation Pipeline

The user interface consists of a web page that allows the user to upload data, input parameters for the pipeline, and include an e-mail address for notification purposes (Figure 5.3.1). Required data input for the pipeline are the FASTQ formatted sequencing files and a library annotation file that links the FASTQ files to sequencing libraries. Since sequencing experiments can be quite large (e.g. the uncompressed human methylome analysis conducted by Ecker's group consumes a few hundred gigabytes per cell line(*1*)), FASTQ files must be compressed before uploaded. Assuming 2 MB/s upload speed, which is easily reached during intercampus transfers, 10 GB can be transferred in 85 minutes. Once the uploaded files have been selected, the user must then input parameters. These parameters include the mapping algorithm preferences, whether to filter for monoclonal reads, and expected gaps between paired ends (if applicable).

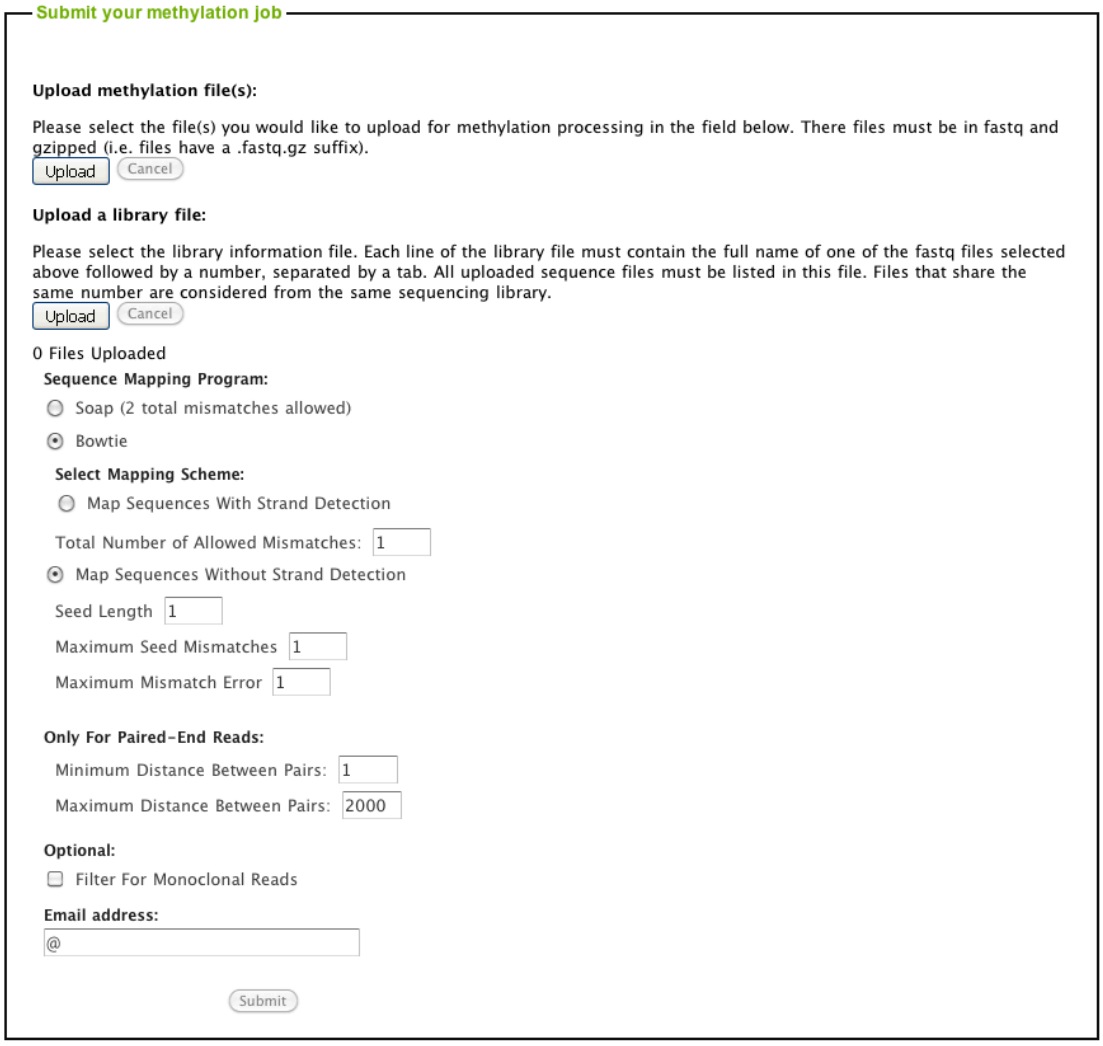


Figure 5.3.1 A screenshot of the methylation pipeline webpage (http://wanglab.ucsd.edu/star/job.php?s=methlyation). This interface allows users to upload FASTQ files for read mapping and methylation-based analyses. Users can select mapping algorithms, enter sequencing end spacing constraints for paired end reads, check monoclonal filtering, and enter his/her e-mail address for job status notification purposes. Users can also upload mapped data and access help files (not shown).

The user also inputs an e-mail address so that he/she receives status updates on his/her job and also a link to download the output from the analysis. The status updates are important so that that user knows whether his/her job is running, queued, or has encountered an error. When a job finishes, there are several files for the user to download: (1) called SNPs, (2) ASM regions, (3) methylation frequencies, (4) LD blocks and, (3) LD Blocks associated with RefSeq genes. Files 2, 3, and 4 are in BED format and are intended for visualization via the UCSC genome browser.

5.4 Integration into the STAR Website

This pipeline is integrated into a next generation sequencing analysis website called STAR (http://wanglab.ucsd.edu/star/). STAR currently processes CHIP-Seq data and RNA-Seq data. The addition of the methylation pipeline adds BS-Seq processing capability. STAR is thusly an All-Seq website that is capable of analyzing various types of sequencing data. Additional work on STAR will involve integrating these various data sources in hopes of improving our understanding of transcriptional regulation. ChIP-Seq data can indicate where transcription factors bind and the presence of histone modifications while BS-Seq data will reveal that methylation states of CpGs. Relationships between methylation, histone modifications, and transcription factors have already been observed (e.g. Cfp1, H3K4me3, and CpG island methylation states as described in Chapter 1). There already exists extensive sequencing data for histone marks for a few human cell lines. IMR90 currently

has ChIP-Seq data for 26 histone marks and this cell line has BS-Seq data available as well, for example. Second generation sequencing technology, such as Illumina sequencing technology, has increased its sequencing capacity greatly in the past couple of years and third generation technology is in development. As the Wei Wang group continually improves STAR, this website will be capable of quickly processing new sequencing data. The group will thusly have the opportunity to get a first chance to find new insights into a cell's regulatory biology.

Chapter 6 Concluding Remarks

The study of DNA methylation via padlock probes has proved useful for studying chromosomal regions of interest. This technique scales easily so that the methylation patterns of particular regions across many different cell types can be investigated. The most recent advancement in targeted bisulfite sequencing is the introduction of the Cpg220k probe set in 2010. An earlier probe set, Cpg30k, demonstrated the capability of this targeted methylation approach to study a subset of CpG islands across 16 cell lines. The generation of these probe sets is dependent on knowing which chromosomal regions are worthy of investigation. Not limited by prior knowledge, shotgun bisulfite sequencing, which is performed in Joe Ecker's group, investigates all CpGs within mappable regions. Both of these analyses produce a lot of data, shotgun bisulfite sequencing for a single cell line and targeted bisulfite sequencing for many cell lines, and an automated pipeline is thusly useful to process this data.

My main contribution to the field of DNA methylation is the creation of tools that process this data. I have established a server based methylation pipeline so that users can simply upload files outputted from Illumina's sequencing software and the pipeline will e-mail them methylation frequencies, called SNPs, SNP-based ASM, and regions of organized methylation patterns. Mapping BSC reads back to the genome is not a trivial process and there are multiple methods that can do this. I implemented two methods, one used in

targeted bisulfite studies and the other used in human methylome studies, to greatly simplify the task of data analysis for the end-user. To facilitate this pipeline, I created a novel SNP calling algorithm for BSC sequences that considers both single and double stranded BSC sequences when calculating genotype probabilities. There are no other programs that do this yet. I then created an algorithm that looks at regions containing SNPs and statistically assessed whether the region is significantly allele specifically methylated. For regions without SNPs, I also applied the LD r^2 algorithm to DNA methylation to identify additional ASM candidates.

In addition to post-experimental data processing, I have also created the new Cpg220k probe set. Using lessons learned from previous probe sets, I created a significantly larger probe set that covered regions predicted to be differentially methylated across various cell types. This involved integrating data from various studies and optimizing the probe set so that the most interesting CpG sites were covered with the least amount of probes (i.e. least cost). Beyond greatly expanded CpG coverage, changes were made to improve the capturing ability of the padlock probes and to capture candidate SNP sites where possible. Through its funding of the Cpg220k probe set, Agilent is exploring the commercial viability of the padlock probe method. When experimental data is produced based on the Cpg220k probe set and processed with my methylation pipeline, I will be involved in both experimental development, via my creation of the Cpg220k probe set, and the analysis of

experimental results. This allows me to take my knowledge based on analyzing previous probe sets and use it to further improve future probe sets.

I believe both general and targeted techniques will be used in the future for studying DNA methylation. BS shotgun sequencing reveals the methylation state of the vast majority of the genome. However, a recent human methylome study shows that most CpGs are methylated and a small percentage of the total CpG population exhibits differential methylation when comparing a differentiated cell to a pluripotent cell(*1*). Targeted sequencing focuses on regions of interest and thusly avoids the costs and effort involved when sequencing an entire genome. However, if no regions of interest are known, the targeted method is useless. I envision large-scale BS shotgun sequencing experiments revealing new CpG sites that undergo biologically relevant methylation changes across conditions. These regions will then be included in targeted probe sets and examined across many different cell lines, which will yield additional biological information about these CpGs.

The integration of these techniques would be useful in identifying CpGs that serve as markers for diseases and diagnosing patients based on the methylation states of these CpGs. For example, a small number of patients exhibiting a diseased phenotype would donate cells that are BS shotgun sequenced. The sequencing data would be analyzed and compared to BS shotgun sequenced data from healthy patients. CpGs whose methylation status discriminated the diseased patients from the healthy patients would then be targeted in a customized padlock probe set. This padlock probe set

would be used on a much larger sample of diseased and healthy patients to test the discriminatory power of the selected CpGs. If these CpGs were shown to be informative in the larger population, the padlock probe set would be used as a clinical diagnostic tool for the general public.

Improvements in sequencing technology will decrease the costs of sequencing and BS shotgun sequencing across many cell lines will thusly become more feasible. Given these advancements, targeted BS sequencing may be seen as technique with a fading purpose. However, these technological advancements will also allow padlock probes to cover more samples. When DNA methylation is used as a diagnostic tool, the methylation states of certain CpGs in thousands to millions of samples will have to be analyzed. Padlock probes are better suited for this task and will thusly remain relevant as technology improves.

Future DNA methylation analyses will focus on allelic biases in DNA methylation as well as overall DNA methylation levels. The ability to discern whether a gene is regulated or expressed monoallelically or biallelically has become increasingly important. Chapter 1 discussed the importance of imprinted genes and the expression of other dosage-specific genes (e.g. MeCP2), which are monoallelically expressed via an ASM regulatory mechanism. As sequencing reads increase in length, the ability to study allele specificity will also improve ASM analyses. Since DNA methylation has been shown to affect chromatin structure (e.g. CTCF-mediated chromatin loops), ASM information can provide further insight in structural differences between

sister chromosomes. By making SNP calls and locating regions of organized methylation, my methylation pipeline will remain relevant as diagnostic tools will likely rely on averaged allelic methylation information while experiments investigating the regulatory effects of DNA methylation will closely examine allele specificity.

References

1. R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, J. R. Ecker, Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315 (2009).

2. R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, J. R. Ecker, Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523 (May 2, 2008).

3. S. J. Cokus, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215 (2008).

4. N. Sasai, M. Nakao, P. A. Defossez, Sequence-specific recognition of methylated DNA by human zinc-finger proteins. *Nucleic Acids Res* **38**, 5015 (Aug 1, 2010).

5. M. Bostick, J. K. Kim, P. O. Esteve, A. Clark, S. Pradhan, S. E. Jacobsen, UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760 (Sep 21, 2007).

6. L. M. Johnson, M. Bostick, X. Zhang, E. Kraft, I. Henderson, J. Callis, S. E. Jacobsen, The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr Biol* **17**, 379 (Feb 20, 2007).

7. G. J. Filion, S. Zhenilo, S. Salozhin, D. Yamada, E. Prokhortchouk, P. A. Defossez, A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol Cell Biol* **26**, 169 (Jan, 2006).

8. J. M. Daniel, C. M. Spring, H. C. Crawford, A. B. Reynolds, A. Baig, The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic Acids Res* **30**, 2911 (Jul 1, 2002).

9. P. A. Defossez, K. F. Kelly, G. J. Filion, R. Perez-Torrado, F. Magdinier, H. Menoni, C. L. Nordgaard, J. M. Daniel, E. Gilson, The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. *J Biol Chem* **280**, 43017 (Dec 30, 2005).

10. J. B. Margot, A. E. Ehrenhofer-Murray, H. Leonhardt, Interactions within the mammalian DNA methyltransferase family. *BMC Mol Biol* **4**, 7 (May 30, 2003).

11. E. Li, Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3**, 662 (Sep, 2002).

12. J. Lan, S. Hua, X. He, Y. Zhang, DNA methyltransferases and methyl-binding proteins of mammals. *Acta Biochim Biophys Sin (Shanghai)* **42**, 243 (Apr, 2010).

13. T. H. Bestor, The DNA methyltransferases of mammals. *Hum Mol Genet* **9**, 2395 (Oct, 2000).

14. M. G. Goll, F. Kirpekar, K. A. Maggert, J. A. Yoder, C. L. Hsieh, X. Zhang, K. G. Golic, S. E. Jacobsen, T. H. Bestor, Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395 (Jan 20, 2006).

15. G. Liang, M. F. Chan, Y. Tomigahara, Y. C. Tsai, F. A. Gonzales, E. Li, P. W. Laird, P. A. Jones, Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Mol Cell Biol* **22**, 480 (Jan, 2002).

16. T. Chen, S. Hevi, F. Gay, N. Tsujimoto, T. He, B. Zhang, Y. Ueda, E. Li, Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. *Nat Genet* **39**, 391 (Mar, 2007).

17. A. Dong, J. A. Yoder, X. Zhang, L. Zhou, T. H. Bestor, X. Cheng, Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. *Nucleic Acids Res* **29**, 439 (Jan 15, 2001).

18. J. A. Yoder, T. H. Bestor, A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. *Hum Mol Genet* **7**, 279 (Feb, 1998).

19. E. Li, T. H. Bestor, R. Jaenisch, Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915 (Jun 12, 1992).

20. H. Gowher, K. Liebert, A. Hermann, G. Xu, A. Jeltsch, Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *J Biol Chem* **280**, 13341 (Apr 8, 2005).

21. Z. X. Chen, J. R. Mann, C. L. Hsieh, A. D. Riggs, F. Chedin, Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. *J Cell Biochem* **95**, 902 (Aug 1, 2005).

22. K. Hata, M. Okano, H. Lei, E. Li, Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983 (Apr, 2002).

23. F. Chedin, M. R. Lieber, C. L. Hsieh, The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc Natl Acad Sci U S A* **99**, 16916 (Dec 24, 2002).

24. C. Holz-Schietinger, N. O. Reich, The inherent processivity of the human de novo DNA methyltransferase 3A (DNMT3A) is enhanced by DNMT3L. *J Biol Chem*, (Jul 14, 2010).

25. J. L. Glass, M. J. Fazzari, A. C. Ferguson-Smith, J. M. Greally, CG dinucleotide periodicities recognized by the Dnmt3a-Dnmt3L complex are distinctive at retroelements and imprinted domains. *Mamm Genome* **20**, 633 (Sep-Oct, 2009).

26. R. Z. Jurkowska, N. Anspach, C. Urbanke, D. Jia, R. Reinhardt, W. Nellen, X. Cheng, A. Jeltsch, Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L. *Nucleic Acids Res* **36**, 6656 (Dec, 2008).

27. S. Feng, S. J. Cokus, X. Zhang, P. Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini, S.

E. Jacobsen, Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**, 8689 (May 11, 2010).

28. A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916 (May 14, 2010).

29. L. M. Field, F. Lyko, M. Mandrioli, G. Prantera, DNA methylation in insects. *Insect Mol Biol* **13**, 109 (Apr, 2004).

30. M. Gold, M. Gefter, R. Hausmann, J. Hurwitz, Methylation of DNA. *J Gen Physiol* **49**, 5 (Jul, 1966).

31. W. Doerfler, Patterns of DNA methylation--evolutionary vestiges of foreign DNA inactivation as a host defense mechanism. A proposal. *Biol Chem Hoppe Seyler* **372**, 557 (Aug, 1991).

32. M. A. Matzke, M. F. Mette, W. Aufsatz, J. Jakowitsch, A. J. Matzke, Host defenses to parasitic sequences and the evolution of epigenetic control mechanisms. *Genetica* **107**, 271 (1999).

33. A. P. Wolffe, M. A. Matzke, Epigenetics: regulation through repression. *Science* **286**, 481 (Oct 15, 1999).

34. D. Zilberman, The evolving functions of DNA methylation. *Curr Opin Plant Biol* **11**, 554 (Oct, 2008).

35. T. Hughes, R. Webb, Y. Fei, J. D. Wren, A. H. Sawalha, DNA methylome in human CD4+ T cells identifies transcriptionally repressive and non-repressive methylation peaks. *Genes Immun*, (May 13, 2010).

36. L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring, C. L. Wei, Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320 (Mar, 2010).

37. H. Xiang, J. Zhu, Q. Chen, F. Dai, X. Li, M. Li, H. Zhang, G. Zhang, D. Li, Y. Dong, L. Zhao, Y. Lin, D. Cheng, J. Yu, J. Sun, X. Zhou, K. Ma, Y. He, Y. Zhao, S. Guo, M. Ye, G. Guo, Y. Li, R. Li, X. Zhang, L. Ma, K. Kristiansen, Q. Guo, J. Jiang, S. Beck, Q. Xia, W. Wang, J. Wang, Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* **28**, 516 (May, 2010).

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

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

40. R. a. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan, A. P. Feinberg,

The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* **41**, 178 (2009).

41. A. Doi, I.-H. Park, B. Wen, P. Murakami, M. J. Aryee, R. Irizarry, B. Herb, C. Ladd-Acosta, J. Rho, S. Loewer, J. Miller, T. Schlaeger, G. Q. Daley, A. P. Feinberg, 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 (2009).

42. M. E. Figueroa, S. Lugthart, Y. Li, C. Erpelinck-Verschueren, X. Deng, P. J. Christos, E. Schifano, J. Booth, W. van Putten, L. Skrabanek, F. Campagne, M. Mazumdar, J. M. Greally, P. J. M. Valk, B. L√∂wenberg, R. Delwel, A. Melnick, DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer cell* **17**, 13 (2010).

43. IARC, *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissue*. E. C. S. Swerdlow, N. Lee Harris, E.S. Jaffe, S.A. Pileri, H. Stein, J. Thiele, J.W. Vardiman, Ed., (World Health Organization, Geneva, ed. 4th, 2008), pp. 441.

44. J. M. Flanagan, S. Cocciardi, N. Waddell, C. N. Johnstone, A. Marsh, S. Henderson, P. Simpson, L. da Silva, K. Khanna, S. Lakhani, C. Boshoff, G. Chenevix-Trench, DNA methylome of familial breast cancer identifies distinct profiles defined by mutation status. *American journal of human genetics* **86**, 420 (2010).

45. T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, A. L. Borresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869 (Sep 11, 2001).

46. A. E. Teschendorff, U. Menon, A. Gentry-Maharaj, S. J. Ramus, S. A. Gayther, S. Apostolidou, A. Jones, M. Lechner, S. Beck, I. J. Jacobs, M. Widschwendter, An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One* **4**, e8274 (2009).

47. R. Grutzmann, B. Molnar, C. Pilarsky, J. K. Habermann, P. M. Schlag, H. D. Saeger, S. Miehlke, T. Stolz, F. Model, U. J. Roblick, H. P. Bruch, R. Koch, V. Liebenberg, T. Devos, X. Song, R. H. Day, A. Z. Sledziewski, C. Lofton-Day, Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS One* **3**, e3759 (2008).

48. T. deVos, R. Tetzner, F. Model, G. Weiss, M. Schuster, J. Distler, K. V. Steiger, R. Grutzmann, C. Pilarsky, J. K. Habermann, P. R. Fleshner, B. M. Oubre, R. Day, A. Z. Sledziewski, C. Lofton-Day, Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clin Chem* **55**, 1337 (Jul, 2009).

49. H. M. Muller, H. Fiegl, A. Widschwendter, M. Widschwendter, Prognostic DNA methylation marker in serum of cancer patients. *Ann N Y Acad Sci* **1022**, 44 (Jun, 2004).

50. H. M. Muller, A. Widschwendter, H. Fiegl, L. Ivarsson, G. Goebel, E. Perkmann, C. Marth, M. Widschwendter, DNA methylation in serum of breast cancer patients: an independent prognostic marker. *Cancer Res* **63**, 7641 (Nov 15, 2003).

51. M. Widschwendter, S. Apostolidou, E. Raum, D. Rothenbacher, H. Fiegl, U. Menon, C. Stegmaier, I. J. Jacobs, H. Brenner, Epigenotyping in peripheral blood cell DNA and breast cancer risk: a proof of principle study. *PLoS One* **3**, e2656 (2008).

52. P. P. Luedi, F. S. Dietrich, J. R. Weidman, J. M. Bosko, R. L. Jirtle, A. J. Hartemink, Computational and experimental identification of novel human imprinted genes. *Genome Res* **17**, 1723 (Dec, 2007).

53. J. M. Frost, G. E. Moore, The importance of imprinting in the human placenta. *Plos Genetics* **6**, e1001015 (2010).

54. Q. J. Hudson, T. M. Kulinski, S. P. Huetter, D. P. Barlow, Genomic imprinting mechanisms in embryonic and extraembryonic mouse tissues. *Heredity* **105**, 45 (Jul, 2010).

55. C. Gregg, J. Zhang, B. Weissbourd, S. Luo, G. P. Schroth, D. Haig, C. Dulac, High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329**, 643 (Aug 6, 2010).

56. K. Kerkel, A. Spadola, E. Yuan, J. Kosek, L. Jiang, E. Hod, K. Li, V. V. Murty, N. Schupf, E. Vilain, Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet* **40**, 904 (2008).

57. C. J. Brown, A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, H. F. Willard, A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38 (Jan 3, 1991).

58. C. J. Brown, R. G. Lafreniere, V. E. Powers, G. Sebastio, A. Ballabio, A. L. Pettigrew, D. H. Ledbetter, E. Levy, I. W. Craig, H. F. Willard, Localization of the X inactivation centre on the human X chromosome in Xq13. *Nature* **349**, 82 (Jan 3, 1991).

59. C. Chureau, M. Prissette, A. Bourdet, V. Barbe, L. Cattolico, L. Jones, A. Eggen, P. Avner, L. Duret, Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res* **12**, 894 (Jun, 2002).

60. H. F. Willard, X chromosome inactivation, XIST, and pursuit of the X-inactivation center. *Cell* **86**, 5 (Jul 12, 1996).

61. J. T. Lee, L. S. Davidow, D. Warshawsky, Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**, 400 (Apr, 1999).

62. J. T. Lee, Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes & development* **23**, 1831 (2009).

63. M. E. Donohoe, S. S. Silva, S. F. Pinter, N. Xu, J. T. Lee, The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature* **460**, 128 (2009).

64. M. C. Anguera, B. K. Sun, N. Xu, J. T. Lee, X-chromosome kiss and tell: how the Xs go their separate ways. *Cold Spring Harb Symp Quant Biol* **71**, 429 (2006).

65. N. Xu, C. L. Tsai, J. T. Lee, Transient homologous chromosome pairing marks the onset of X inactivation. *Science* **311**, 1149 (Feb 24, 2006).

66. J. T. Lee, Molecular links between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting? *Curr Biol* **13**, R242 (Mar 18, 2003).

67. N. Xu, M. E. Donohoe, S. S. Silva, J. T. Lee, Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. *Nat Genet* **39**, 1390 (Nov, 2007).

68. T. S. Barakat, I. Jonkers, K. Monkhorst, J. Gribnau, X-changing information on X inactivation. *Exp Cell Res* **316**, 679 (Mar 10, 2010).

69. I. Jonkers, T. S. Barakat, E. M. Achame, K. Monkhorst, A. Kenter, E. Rentmeester, F. Grosveld, J. A. Grootegoed, J. Gribnau, RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation. *Cell* **139**, 999 (Nov 25, 2009).

70. B. K. Sun, A. M. Deaton, J. T. Lee, A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. *Mol Cell* **21**, 617 (Mar 3, 2006).

71. J. Zhao, B. K. Sun, J. A. Erwin, J. J. Song, J. T. Lee, Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750 (Oct 31, 2008).

72. S. Maenner, M. Blaud, L. Fouillen, A. Savoye, V. Marchand, A. Dubois, S. Sanglier-Cianferani, A. Van Dorsselaer, P. Clerc, P. Avner, A. Visvikis, C. Branlant, 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8**, e1000276 (Jan, 2010).

73. A. Wutz, T. P. Rasmussen, R. Jaenisch, Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet* **30**, 167 (Feb, 2002).

74. E. Vire, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J. M. Vanderwinden, M. Bollen, M. Esteller, L. Di Croce, Y. de Launoit, F. Fuks, The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871 (Feb 16, 2006).

75. D. Zilberman, X. Cao, L. K. Johansen, Z. Xie, J. C. Carrington, S. E. Jacobsen, Role of Arabidopsis ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr Biol* **14**, 1214 (Jul 13, 2004).

76. Y. Qi, X. He, X. J. Wang, O. Kohany, J. Jurka, G. J. Hannon, Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443**, 1008 (Oct 26, 2006).

77. Z. Gao, H. L. Liu, L. Daxinger, O. Pontes, X. He, W. Qian, H. Lin, M. Xie, Z. J. Lorkovic, S. Zhang, D. Miki, X. Zhan, D. Pontier, T. Lagrange, H. Jin, A. J. Matzke, M. Matzke, C. S. Pikaard, J. K. Zhu, An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature* **465**, 106 (May 6, 2010).

78. X. J. He, Y. F. Hsu, S. Zhu, A. T. Wierzbicki, O. Pontes, C. S. Pikaard, H. L. Liu, C. S. Wang, H. Jin, J. K. Zhu, An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell* **137**, 498 (May 1, 2009).

79. D. Lucifero, C. Mertineit, H. J. Clarke, T. H. Bestor, J. M. Trasler, Methylation dynamics of imprinted genes in mouse germ cells. *Genomics* **79**, 530 (Apr, 2002).

80. J. Y. Li, D. J. Lees-Murdock, G. L. Xu, C. P. Walsh, Timing of establishment of paternal methylation imprints in the mouse. *Genomics* **84**, 952 (Dec, 2004).

81. T. L. Davis, J. M. Trasler, S. B. Moss, G. J. Yang, M. S. Bartolomei, Acquisition of the H19 methylation imprint occurs differentially on the parental alleles during spermatogenesis. *Genomics* **58**, 18 (May 15, 1999).

82. D. Bourc'his, C. Proudhon, Sexual dimorphism in parental imprint ontogeny and contribution to embryonic development. *Mol Cell Endocrinol* **282**, 87 (Jan 30, 2008).

83. R. Hirasawa, H. Chiba, M. Kaneda, S. Tajima, E. Li, R. Jaenisch, H. Sasaki, Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev* **22**, 1607 (Jun 15, 2008).

84. Q. Zhao, G. Rank, Y. T. Tan, H. Li, R. L. Moritz, R. J. Simpson, L. Cerruti, D. J. Curtis, D. J. Patel, C. D. Allis, J. M. Cunningham, S. M. Jane, PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nat Struct Mol Biol* **16**, 304 (Mar, 2009).

85. D. Jia, R. Z. Jurkowska, X. Zhang, A. Jeltsch, X. Cheng, Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**, 248 (Sep 13, 2007).

86. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, R. A. Young, A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77 (Jul 13, 2007).

87. N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, B. Ren, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311 (Mar, 2007).

88. M. V. Koerner, F. M. Pauler, R. Huang, D. P. Barlow, The function of non-coding RNAs in genomic imprinting. *Development* **136**, 1771 (Jun, 2009).

89. D. Mancini-Dinardo, S. J. Steele, J. M. Levorse, R. S. Ingram, S. M. Tilghman, Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20**, 1268 (May 15, 2006).

90. J. Y. Shin, G. V. Fitzpatrick, M. J. Higgins, Two distinct mechanisms of silencing by the KvDMR1 imprinting control region. *EMBO J* **27**, 168 (Jan 9, 2008).

91. T. Sahoo, D. del Gaudio, J. R. German, M. Shinawi, S. U. Peters, R. E. Person, A. Garnica, S. W. Cheung, A. L. Beaudet, Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet* **40**, 719 (Jun, 2008).

92. M. Bastepe, J. E. Pincus, T. Sugimoto, K. Tojo, M. Kanatani, Y. Azuma, K. Kruse, a. L. Rosenbloom, H. Koshiyama, H. J√ºppner, Positional dissociation between the genetic mutation responsible for pseudohypoparathyroidism type Ib and the associated methylation defect at exon A/B: evidence for a long-range regulatory

element within the imprinted GNAS1 locus. *Human molecular genetics* **10**, 1231 (2001).

93. B. Lecumberri, E. Fern√°ndez-Rebollo, L. Sentchordi, P. Saavedra, a. Bernal-Chico, L. F. Pallardo, J. M. J. n. Bustos, L. Casta√±o, M. de Santiago, O. Hiort, G., M. Bastepe, Coexistence of two different pseudohypoparathyroidism subtypes (Ia and Ib) in the same kindred with independent Gs{alpha} coding mutations and GNAS imprinting defects. *Journal of medical genetics* **47**, 276 (2010).

94. F. Cerrato, A. Sparago, G. Verde, A. D. Crescenzo, V. Citro, M. V. Cubellis, M. M. Rinaldi, L. Boccuto, G. Neri, C. Magnani, P. D‚Äôangelo, P. Collini, D. Perotti, G. Sebastio, E. R. Maher, A. Riccio, Different mechanisms cause imprinting defects at the IGF2/H19 locus in Beckwith-Wiedemann syndrome and Wilms‚Äô tumour, Hum. *Mol. Genet.* **17**, 1427 (2008).

95. P. J. Skene, R. S. Illingworth, S. Webb, A. R. W. Kerr, K. D. James, D. J. Turner, R. Andrews, A. P. Bird, Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Molecular cell* **37**, 457 (2010).

96. E. Schilling, C. El Chartouni, M. Rehli, Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. *Genome Res* **19**, 2028 (Nov, 2009).

97. A. Murrell, S. Heeson, W. N. Cooper, E. Douglas, S. Apostolidou, G. E. Moore, E. R. Maher, W. Reik, An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype. *Hum Mol Genet* **13**, 247 (Jan 15, 2004).

98. J. M. Flanagan, V. Popendikyte, N. Pozdniakovaite, M. Sobolev, A. Assadzadeh, A. Schumacher, M. Zangeneh, L. Lau, C. Virtanen, S. C. Wang, A. Petronis, Intra- and interindividual epigenetic variation in human germ cells. *Am J Hum Genet* **79**, 67 (Jul, 2006).

99. B. T. Heijmans, D. Kremer, E. W. Tobi, D. I. Boomsma, P. E. Slagboom, Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum Mol Genet* **16**, 547 (Mar 1, 2007).

100. L. C. Schalkwyk, E. L. Meaburn, R. Smith, E. L. Dempster, A. R. Jeffries, M. N. Davies, R. Plomin, J. Mill, Allelic skewing of DNA methylation is widespread across the genome. *Am J Hum Genet* **86**, 196 (Feb 12, 2010).

101. V. Plagnol, E. Uz, C. Wallace, H. Stevens, D. Clayton, T. Ozcelik, J. A. Todd, Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS One* **3**, e2966 (2008).

102. E. Choy, R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, M. Rivas, E. T. Dermitzakis, E. Cahir-McFarland, E. Kieff, D. Hafler, M. J. Daly, D. Altshuler, Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *Plos Genet* **4**, e1000287 (Nov, 2008).

103. J. H. Lee, I. H. Park, Y. Gao, J. B. Li, Z. Li, G. Q. Daley, K. Zhang, G. M. Church, A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *Plos Genetics* **5**, e1000718 (Nov, 2009).

104. A. Gimelbrant, J. N. Hutchinson, B. R. Thompson, A. Chess, Widespread monoallelic expression on human autosomes. *Science* **318**, 1136 (Nov 16, 2007).

105. K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663 (Aug 25, 2006).

106. J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, Slukvin, II, J. A. Thomson, Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917 (Dec 21, 2007).

107. J. Yu, M. A. Vodyanik, P. He, Slukvin, II, J. A. Thomson, Human embryonic stem cells reprogram myeloid precursors following cell-cell fusion. *Stem Cells* **24**, 168 (Jan, 2006).

108. J. Yu, K. Hu, K. Smuga-Otto, S. Tian, R. Stewart, Slukvin, II, J. A. Thomson, Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324**, 797 (May 8, 2009).

109. K. Okita, T. Ichisaka, S. Yamanaka, Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313 (Jul 19, 2007).

110. A. Banito, J. Gil, Induced pluripotent stem cells and senescence: learning the biology to improve the technology. *EMBO Rep* **11**, 353 (May, 2010).

111. K. Kaji, K. Norrby, A. Paca, M. Mileikovsky, P. Mohseni, K. Woltjen, Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* **458**, 771 (Apr 9, 2009).

112. T. Kawamura, J. Suzuki, Y. V. Wang, S. Menendez, L. B. Morera, A. Raya, G. M. Wahl, J. C. Belmonte, Linking the p53 tumour suppressor pathway to somatic cell reprogramming. *Nature* **460**, 1140 (Aug 27, 2009).

113. K. Woltjen, I. P. Michael, P. Mohseni, R. Desai, M. Mileikovsky, R. Hamalainen, R. Cowling, W. Wang, P. Liu, M. Gertsenstein, K. Kaji, H. K. Sung, A. Nagy, piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458**, 766 (Apr 9, 2009).

114. K. Okita, M. Nakagawa, H. Hyenjong, T. Ichisaka, S. Yamanaka, Generation of mouse induced pluripotent stem cells without viral vectors. *Science* **322**, 949 (Nov 7, 2008).

115. F. Jia, K. D. Wilson, N. Sun, D. M. Gupta, M. Huang, Z. Li, N. J. Panetta, Z. Y. Chen, R. C. Robbins, M. A. Kay, M. T. Longaker, J. C. Wu, A nonviral minicircle vector for deriving human iPS cells. *Nat Methods* **7**, 197 (Mar, 2010).

116. A. Longacre, U. Storb, A novel cytidine deaminase affects antibody diversity. *Cell* **102**, 541 (Sep 1, 2000).

117. M. Muramatsu, K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, T. Honjo, Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553 (Sep 1, 2000).

118. P. Revy, T. Muto, Y. Levy, F. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Dufourcq-Labelouse, A. Gennery, I. Tezcan, F. Ersoy, H. Kayserili, A. G. Ugazio, N. Brousse, M. Muramatsu, L. D. Notarangelo, K. Kinoshita, T. Honjo, A. Fischer, A. Durandy, Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* **102**, 565 (Sep 1, 2000).

119. H. D. Morgan, W. Dean, H. A. Coker, W. Reik, S. K. Petersen-Mahrt, Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J Biol Chem* **279**, 52353 (Dec 10, 2004).

120. S. G. Conticello, M. A. Langlois, Z. Yang, M. S. Neuberger, DNA deamination in immunity: AID in the context of its APOBEC relatives. *Adv Immunol* **94**, 37 (2007).

121. D. A. Macduff, R. S. Harris, Directed DNA deamination by AID/APOBEC3 in immunity. *Curr Biol* **16**, R186 (Mar 21, 2006).

122. K. Rai, I. J. Huggins, S. R. James, A. R. Karpf, D. A. Jones, B. R. Cairns, DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell* **135**, 1201 (Dec 26, 2008).

123. B. Hendrich, U. Hardeland, H. H. Ng, J. Jiricny, A. Bird, The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301 (Sep 16, 1999).

124. N. Bhutani, J. J. Brady, M. Damian, A. Sacco, S. p. Y. Corbel, H. M. Blau, Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* **463**, 1042 (2010).

125. H. Sasaki, Y. Matsui, Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat Rev Genet* **9**, 129 (Feb, 2008).

126. P. Hajkova, K. Ancelin, T. Waldmann, N. Lacoste, U. C. Lange, F. Cesari, C. Lee, G. Almouzni, R. Schneider, M. A. Surani, Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature* **452**, 877 (Apr 17, 2008).

127. W. Reik, W. Dean, J. Walter, Epigenetic reprogramming in mammalian development. *Science* **293**, 1089 (Aug 10, 2001).

128. W. Mayer, A. Niveleau, J. Walter, R. Fundele, T. Haaf, Demethylation of the zygotic paternal genome. *Nature* **403**, 501 (Feb 3, 2000).

129. P. Hajkova, S. Erhardt, N. Lane, T. Haaf, O. El-Maarri, W. Reik, J. Walter, M. A. Surani, Epigenetic reprogramming in mouse primordial germ cells. *Mech Dev* **117**, 15 (Sep, 2002).

130. C. Popp, W. Dean, S. Feng, S. J. Cokus, S. Andrews, M. Pellegrini, S. E. Jacobsen, W. Reik, Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101 (2010).

131. Y. Seki, M. Yamaji, Y. Yabuta, M. Sano, M. Shigeta, Y. Matsui, Y. Saga, M. Tachibana, Y. Shinkai, M. Saitou, Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. *Development* **134**, 2627 (Jul, 2007).

132. S. Daujat, T. Weiss, F. Mohn, U. C. Lange, C. Ziegler-Birling, U. Zeissler, M. Lappe, D. Schubeler, M. E. Torres-Padilla, R. Schneider, H3K64 trimethylation marks heterochromatin and is dynamically remodeled during developmental reprogramming. *Nat Struct Mol Biol* **16**, 777 (Jul, 2009).

133. S. M. Chuva de Sousa Lopes, K. Hayashi, T. C. Shovlin, W. Mifsud, M. A. Surani, A. McLaren, X chromosome activity in mouse XX primordial germ cells. *PLoS Genet* **4**, e30 (Feb, 2008).

134. J. P. Thomson, P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. W. Kerr, A. e. Deaton, R. Andrews, K. D. James, D. J. Turner, R. Illingworth, A. Bird, CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082 (2010).

135. J. H. Lee, C. M. Tate, J. S. You, D. G. Skalnik, Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *J Biol Chem* **282**, 13419 (May 4, 2007).

136. J. H. Lee, D. G. Skalnik, CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem* **280**, 41725 (Dec 16, 2005).

137. X. Xie, T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, E. S. Lander, Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**, 7145 (Apr 24, 2007).

138. T. Li, J. F. Hu, X. Qiu, J. Ling, H. Chen, S. Wang, A. Hou, T. H. Vu, A. R. Hoffman, CTCF regulates allelic expression of Igf2 by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loop. *Mol Cell Biol* **28**, 6473 (2008).

139. S. Cuddapah, R. Jothi, D. E. Schones, T. Y. Roh, K. Cui, K. Zhao, Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24 (Jan, 2009).

140. W. M. Gombert, A. Krumm, Targeted deletion of multiple CTCF-binding elements in the human C-MYC gene reveals a requirement for CTCF in C-MYC expression. *PloS one* **4**, e6109 (2009).

141. K. Ishihara, M. Oshimura, M. Nakao, CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell* **23**, 733 (Sep 1, 2006).

142. M. Chahrour, S. Y. Jung, C. Shaw, X. Zhou, S. T. Wong, J. Qin, H. Y. Zoghbi, MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science* **320**, 1224 (May 30, 2008).

143. Y. a. Boumber, Y. Kondo, X. Chen, L. Shen, Y. Guo, C. Tellez, M. R. H. Est√©cio, S. Ahmed, J.-P. J. Issa, An Sp1/Sp3 binding polymorphism confers methylation protection. *PLoS genetics* **4**, e1000162 (2008).

144. M. Brandeis, D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn, A. Nemes, V. Temper, A. Razin, H. Cedar, Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**, 435 (Sep 29, 1994).

145. R. Ohlsson, R. Renkawitz, V. Lobanenkov, CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* **17**, 520 (Sep, 2001).

146. J. E. Phillips, V. G. Corces, CTCF: master weaver of the genome. *Cell* **137**, 1194 (Jun 26, 2009).

147. A. Baniahmad, C. Steiner, A. C. Kohne, R. Renkawitz, Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. *Cell* **61**, 505 (May 4, 1990).

148. E. M. Klenova, R. H. Nicolas, H. F. Paterson, A. F. Carne, C. M. Heath, G. H. Goodwin, P. E. Neiman, V. V. Lobanenkov, CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol Cell Biol* **13**, 7612 (Dec, 1993).

149. V. V. Lobanenkov, R. H. Nicolas, V. V. Adler, H. Paterson, E. M. Klenova, A. V. Polotskaja, G. H. Goodwin, A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* **5**, 1743 (Dec, 1990).

150. A. A. Vostrov, W. W. Quitschke, The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J Biol Chem* **272**, 33353 (Dec 26, 1997).

151. A. C. Bell, G. Felsenfeld, Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482 (May 25, 2000).

152. A. T. Hark, C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, S. M. Tilghman, CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486 (May 25, 2000).

153. P. Szabo, S. H. Tang, A. Rentsendorj, G. P. Pfeifer, J. R. Mann, Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. *Curr Biol* **10**, 607 (May 18, 2000).

154. Y. S. Yoon, S. Jeong, Q. Rong, K. Y. Park, J. H. Chung, K. Pfeifer, Analysis of the H19ICR insulator. *Mol Cell Biol* **27**, 3499 (May, 2007).

155. I. A. De La Rosa-Velazquez, H. Rincon-Arano, L. Benitez-Bribiesca, F. Recillas-Targa, Epigenetic regulation of the human retinoblastoma tumor suppressor gene promoter by CTCF. *Cancer Res* **67**, 2577 (Mar 15, 2007).

156. K. S. Voo, D. L. Carlone, B. M. Jacobsen, A. Flodin, D. G. Skalnik, Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol* **20**, 2108 (Mar, 2000).

157. J. H. Lee, K. S. Voo, D. G. Skalnik, Identification and characterization of the DNA binding domain of CpG-binding protein. *J Biol Chem* **276**, 44669 (Nov 30, 2001).

158. X. Shi, I. Kachirskaia, K. L. Walter, J. H. Kuo, A. Lake, F. Davrazou, S. M. Chan, D. G. Martin, I. M. Fingerman, S. D. Briggs, L. Howe, P. J. Utz, T. G. Kutateladze, A. A. Lugovskoy, M. T. Bedford, O. Gozani, Proteome-wide analysis in Saccharomyces cerevisiae identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *J Biol Chem* **282**, 2450 (Jan 26, 2007).

159. T. S. Mikkelsen, J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, A. Meissner, Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49 (Jul 3, 2008).

160. R. R. Meehan, J. D. Lewis, S. McKay, E. L. Kleiner, A. P. Bird, Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**, 499 (Aug 11, 1989).

161. X. Nan, F. J. Campoy, A. Bird, MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88**, 471 (Feb 21, 1997).

162. O. Bogdanovic, G. J. Veenstra, DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma* **118**, 549 (Oct, 2009).

163. T. Clouaire, I. Stancheva, Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? *Cell Mol Life Sci* **65**, 1509 (May, 2008).

164. B. Hendrich, S. Tweedie, The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* **19**, 269 (May, 2003).

165. Y. Zhang, H. H. Ng, H. Erdjument-Bromage, P. Tempst, A. Bird, D. Reinberg, Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev* **13**, 1924 (Aug 1, 1999).

166. F. Fuks, P. J. Hurd, D. Wolf, X. Nan, A. P. Bird, T. Kouzarides, The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J Biol Chem* **278**, 4035 (Feb 7, 2003).

167. X. Nan, H. H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman, A. Bird, Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386 (May 28, 1998).

168. P. L. Jones, G. J. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis, A. P. Wolffe, Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* **19**, 187 (Jun, 1998).

169. D. Balmer, J. Goldstine, Y. M. Rao, J. M. LaSalle, Elevated methyl-CpG-binding protein 2 expression is acquired during postnatal human brain development and is correlated with alternative polyadenylation. *J Mol Med* **81**, 61 (Jan, 2003).

170. M. L. Gonzales, J. M. LaSalle, The role of MeCP2 in brain development and neurodevelopmental disorders. *Curr Psychiatry Rep* **12**, 127 (Apr, 2010).

171. J. Guy, B. Hendrich, M. Holmes, J. E. Martin, A. Bird, A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. *Nat Genet* **27**, 322 (Mar, 2001).

172. A. L. Collins, J. M. Levenson, A. P. Vilaythong, R. Richman, D. L. Armstrong, J. L. Noebels, J. David Sweatt, H. Y. Zoghbi, Mild overexpression of MeCP2 causes a progressive neurological disorder in mice. *Hum Mol Genet* **13**, 2679 (Nov 1, 2004).

173. C. Gebhard, C. Benner, M. Ehrich, L. Schwarzfischer, E. Schilling, M. Klug, W. Dietmaier, C. Thiede, E. Holler, R. Andreesen, M. Rehli, General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer Res* **70**, 1398 (Feb 15, 2010).

174. G. Caretti, M. Di Padova, B. Micales, G. E. Lyons, V. Sartorelli, The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. *Genes Dev* **18**, 2627 (Nov 1, 2004).

175. A. M. Lindroth, Y. J. Park, C. M. McLean, G. A. Dokshin, J. M. Persson, H. Herman, D. Pasini, X. Miro, M. E. Donohoe, J. T. Lee, K. Helin, P. D. Soloway, Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS Genet* **4**, e1000145 (2008).

176. S. Kriaucionis, N. Heintz, The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (New York, N.Y.)* **324**, 929 (2009).

177. M. Tahiliani, K. P. Koh, Y. Shen, W. a. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, A. Rao, Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)* **324**, 930 (2009).

178. L. J. Cliffe, R. Kieft, T. Southern, S. R. Birkeland, M. Marshall, K. Sweeney, R. Sabatini, JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res* **37**, 1452 (Apr, 2009).

179. M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, D. Schubeler, Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853 (Aug, 2005).

180. I. M. Wilson, J. J. Davies, M. Weber, C. J. Brown, C. E. Alvarez, C. MacAulay, D. Schubeler, W. L. Lam, Epigenomics: mapping the methylome. *Cell Cycle* **5**, 155 (Jan, 2006).

181. F. V. Jacinto, E. Ballestar, M. Esteller, Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques* **44**, 35 (Jan, 2008).

182. M. F. Fraga, E. Ballestar, G. Montoya, P. Taysavang, P. A. Wade, M. Esteller, The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res* **31**, 1765 (Mar 15, 2003).

183. M. Fatemi, P. A. Wade, MBD family proteins: reading the epigenetic code. *J Cell Sci* **119**, 3033 (Aug 1, 2006).

184. N. Sasai, P. A. Defossez, Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes. *Int J Dev Biol* **53**, 323 (2009).

185. B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach, S. W. Turner, Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461 (Jun, 2010).

186. R. a. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. a. Brandenburg, J. a. Jeddeloh, B. Wen, A. P. Feinberg, Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome research* **18**, 780 (2008).

187. M. Oda, J. L. Glass, R. F. Thompson, Y. Mo, E. N. Olivier, M. E. Figueroa, R. R. Selzer, T. A. Richmond, X. Zhang, L. Dannenberg, R. D. Green, A. Melnick, E. Hatchwell, E. E. Bouhassira, A. Verma, M. Suzuki, J. M. Greally, High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res* **37**, 3829 (Jul, 2009).

188. B. Khulan, R. F. Thompson, K. Ye, M. J. Fazzari, M. Suzuki, E. Stasiek, M. E. Figueroa, J. L. Glass, Q. Chen, C. Montagna, E. Hatchwell, R. R. Selzer, T. A. Richmond, R. D. Green, A. Melnick, J. M. Greally, Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res* **16**, 1046 (Aug, 2006).

189. S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, S. E. Jacobsen, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215 (Mar 13, 2008).

190. M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, C. L. Paul, A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827 (Mar 1, 1992).

191. M. P. Ball, J. B. Li, Y. Gao, J.-H. Lee, E. M. LeProust, I.-H. Park, B. Xie, G. Q. Daley, G. M. Church, Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27**, 361 (2009).

192. G. Felsenfeld, M. Groudine, Controlling the double helix. *Nature* **421**, 448 (Jan 23, 2003).

193. N. J. Sakabe, M. A. Nobrega, Genome-wide maps of transcription regulatory elements. *Wiley Interdiscip Rev Syst Biol Med* **2**, 422 (Jul, 2010).

194. D. R. Higgs, D. Vernimmen, J. Hughes, R. Gibbons, Using genomics to study how chromatin influences gene expression. *Annu Rev Genomics Hum Genet* **8**, 299 (2007).

195. N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, B. Ren, Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108 (May 7, 2009).

196. K. J. Won, I. Chepelev, B. Ren, W. Wang, Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* **9**, 547 (2008).

197. K. J. Won, S. Agarwal, L. Shen, R. Shoemaker, B. Ren, W. Wang, An integrated approach to identifying cis-regulatory modules in the human genome. *PLoS One* **4**, e5501 (2009).

198. M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, R. A. Young, Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, (Aug 18, 2010).

199. J. R. Schoenborn, M. O. Dorschner, M. Sekimata, D. M. Santer, M. Shnyreva, D. R. Fitzpatrick, J. A. Stamatoyannopoulos, C. B. Wilson, Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding interferon-gamma. *Nat Immunol* **8**, 732 (Jul, 2007).

200. P. J. Sabo, M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble, J. A. Stamatoyannopoulos, Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3**, 511 (Jul, 2006).

201. J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, J. A. Stamatoyannopoulos, Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**, 283 (Apr, 2009).

202. N. Gheldof, E. M. Smith, T. M. Tabuchi, C. M. Koch, I. Dunham, J. A. Stamatoyannopoulos, J. Dekker, Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic Acids Res* **38**, 4325 (Jul, 2010).

203. V. Parelho, S. Hadjur, M. Spivakov, M. Leleu, S. Sauer, H. C. Gregson, A. Jarmuz, C. Canzonetta, Z. Webster, T. Nesterova, B. S. Cobb, K. Yokomori, N. Dillon, L. Aragon, A. G. Fisher, M. Merkenschlager, Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422 (Feb 8, 2008).

204. R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, J. A. Stamatoyannopoulos, Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**, 139 (Jan 5, 2010).

205. N. R. Markham, M. Zuker, UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**, 3 (2008).

206. R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966 (Aug 1, 2009).

207. I. H. Park, R. Zhao, J. A. West, A. Yabuuchi, H. Huo, T. A. Ince, P. H. Lerou, M. W. Lensch, G. Q. Daley, Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141 (Jan 10, 2008).

208. N. Maherali, T. Ahfeldt, A. Rigamonti, J. Utikal, C. Cowan, K. Hochedlinger, A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell* **3**, 340 (Sep 11, 2008).

209. C. A. Cowan, I. Klimanskaya, J. McMahon, J. Atienza, J. Witmyer, J. P. Zucker, S. Wang, C. C. Morton, A. P. McMahon, D. Powers, D. A. Melton, Derivation of embryonic stem-cell lines from human blastocysts. *N Engl J Med* **350**, 1353 (Mar 25, 2004).

210. A. E. Chen, D. Egli, K. Niakan, J. Deng, H. Akutsu, M. Yamaki, C. Cowan, C. Fitz-Gerald, K. Zhang, D. A. Melton, K. Eggan, Optimal timing of inner cell mass isolation increases the efficiency of human embryonic stem cell derivation and allows generation of sibling cell lines. *Cell Stem Cell* **4**, 103 (Feb 6, 2009).

211. K. Osafune, L. Caron, M. Borowiak, R. J. Martinez, C. S. Fitz-Gerald, Y. Sato, C. A. Cowan, K. R. Chien, D. A. Melton, Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat Biotechnol* **26**, 313 (Mar, 2008).

212. T. Eggermann, Silver-Russell and Beckwith-Wiedemann syndromes: opposite (epi)mutations in 11p15 result in opposite clinical pictures. *Horm Res* **71 Suppl 2**, 30 (Apr, 2009).

213. K. D. Robertson, DNA methylation and human disease. *Nat Rev Genet* **6**, 597 (Aug, 2005).

214. K. Kerkel, A. Spadola, E. Yuan, J. Kosek, L. Jiang, E. Hod, K. Li, V. V. Murty, N. Schupf, E. Vilain, M. Morris, F. Haghighi, B. Tycko, Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* **40**, 904 (Jul, 2008).

215. Y. Zhang, C. Rohde, R. Reinhardt, C. Voelcker-Rehage, A. Jeltsch, Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biol* **10**, R138 (2009).

216. M. Spivakov, A. G. Fisher, Epigenetic signatures of stem-cell identity. *Nat Rev Genet* **8**, 263 (Apr, 2007).

217. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (Aug 15, 2009).

218. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K.

P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich *et al.*, Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78 (Jan 1).

219. K. Zhang, J. B. Li, Y. Gao, D. Egli, B. Xie, J. Deng, Z. Li, J. H. Lee, J. Aach, E. M. Leproust, K. Eggan, G. M. Church, Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6**, 613 (Aug, 2009).

220. B. Ge, D. K. Pokholok, T. Kwan, E. Grundberg, L. Morcos, D. J. Verlaan, J. Le, V. Koka, K. C. Lam, V. Gagne, J. Dias, R. Hoberman, A. Montpetit, M. M. Joly, E. J. Harvey, D. Sinnett, P. Beaulieu, R. Hamon, A. Graziani, K. Dewar, E. Harmsen, J. Majewski, H. H. Goring, A. K. Naumova, M. Blanchette, K. L. Gunderson, T. Pastinen, Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat Genet*, (Oct 18, 2009).

221. G. A. Heap, J. H. Yang, K. Downes, B. C. Healy, K. A. Hunt, N. Bockett, L. Franke, P. C. Dubois, C. A. Mein, R. J. Dobson, T. J. Albert, M. J. Rodesch, D. G. Clayton, J. A. Todd, D. A. van Heel, V. Plagnol, Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. *Hum Mol Genet*, (Oct 13, 2009).

222. H. Yan, W. Yuan, V. E. Velculescu, B. Vogelstein, K. W. Kinzler, Allelic variation in human gene expression. *Science* **297**, 1143 (Aug 16, 2002).

223. J. H. Lee, I. H. Park, Y. Gao, J. B. Li, Z. Li, G. Q. Daley, K. Zhang, G. M. Church, A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* **5**, e1000718 (Nov, 2009).

224. T. Pastinen, T. J. Hudson, Cis-acting regulatory variation in the human genome. *Science* **306**, 647 (Oct 22, 2004).

225. R. Ghotbi, A. Gomez, L. Milani, G. Tybring, A. C. Syvanen, L. Bertilsson, M. Ingelman-Sundberg, E. Aklillu, Allele-specific expression and gene methylation in the control of CYP1A2 mRNA level in human livers. *Pharmacogenomics J* **9**, 208 (Jun, 2009).

226. L. Milani, A. Lundmark, J. Nordlund, A. Kiialainen, T. Flaegstad, G. Jonmundsson, J. Kanerva, K. Schmiegelow, K. L. Gunderson, G. Lonnerholm, A. C. Syvanen, Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Res* **19**, 1 (Jan, 2009).

227. J. B. Li, Y. Gao, J. Aach, K. Zhang, G. V. Kryukov, B. Xie, A. Ahlford, J. K. Yoon, A. M. Rosenbaum, A. W. Zaranek, E. LeProust, S. R. Sunyaev, G. M. Church, Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res* **19**, 1606 (Sep, 2009).

228. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

229. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (Jul 15, 2009).

230. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656 (Apr, 2002).