THE EVOLUTIONARY SIGNIFICANCE OF

DNA METHYLATION IN HUMAN GENOME

A Dissertation
Presented to
The Academic Faculty

by

Jia Zeng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biology

Georgia Institute of Technology
December 2013

THE EVOLUTIONARY SIGNIFICANCE OF

DNA METHYLATION IN HUMAN GENOME

Approved by:

Dr. Soojin Yi, Advisor
School of Biology
Georgia Institute of Technology

Dr. Todd Streelman
School of Biology
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Todd Preuss
Yerkes National Primate Research Center
Emory University

Dr. Michael Goodisman
School of Biology
Georgia Institute of Technology

To my family, friends and advisors

ACKNOWLEDGEMENTS

Pursuing the degree of Ph.D. is a long journey for me with pain, gain tear and laugh. I am sure I can't make it without the love and support of colleagues, family and friends. First of all, my parents deserve much credit for their patience and effort in raising me up, and also for their faith that this day would come one day. There is a saying in China: If your parents are still alive, you shouldn't be far away from home. Even though I could do good work and research during my Ph.D., I never could be a good son when I am ten thousand kilometers away from home for more than five years. I do not think I will ever be able to thank them though and compensate for what they have done for me. I would not be where I am today without their unconditional love and support.

It is my best luck in my life that I have been to have the unconditional love of support from my wife, Minmin. She is always willing to take off and share my burdens. She understood when I have to work overnight to meet the deadline, believed in me when I was in the bottom of my life. Without her, the past five years would be a lonely, suffering and painful time period for me. I am so grateful for her accompany and support for me. Her unselfish support of my dreams and ambitions can't be exaggerated, and I can only hope that I am able to compensate her over time.

Professionally, this work would not have been possible without the efforts of my advisor, Dr Soojin Yi. Her tireless writing of proposals provided me with the opportunity to conduct research without significant concern over lost funding or lack of supplies. She is one of the hardest working people in the department, and I never once regretted joining

the group. I have been stayed in school for more than twenty years and got in touch with many teachers and mentors. Soojin is the closest one to me in my life. What is more important, her attitude, thought and value of science must affect and guide the rest of my career. I would also like to thank you Dr King Jordan, Dr. Michael Goodisman, Dr. Todd Streelman and Dr. Todd Preuss for taking the time out to guide my research and provide me feedback and advice. I also must thank Dr. Brendan Hunt and Dr. Eddie Loh. As senior students in the lab, they took good care of me and helped me a lot in teaching me how to perform scientific analysis as well as think more independently, for their brutal honesty, friendship, quick-wit and uncanny attention to detail. They were the tremendous role models for me during my time at Georgia Tech. I would especially would like to thank my collaborators Dr. Richard Clark from University of Utah, Dr Genevieve Konopka from UCLA, Dr. Taesung Park, Jungsun Park and Iksoo Huh from Seoul National University. Without their hard work, there won't be those publications where I am the author and co-author. At last, I also must thank all my current and previous lab mates Dr. Zuogang Peng, Dr. Thomas Keller, Dr. Ke Xu, Shrutii Sarda, Hema Nagrajan and Dan Sun for their technical expertise, scientific advice and willingness to discuss ideas.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

SUMMARY

In eukaryotic genomes ranging from plants to mammals, DNA methylation is a major epigenetic modification of DNA by adding a methyl group exclusively to cytosine residuals. In mammalian genomes such as humans, these cytosine bases are usually followed by guanine [1]. Although it does not change the primary DNA sequence, this covalent modification plays critical roles in several regulatory processes and can impact gene activity in a heritable fashion [2, 3]. What is more important, DNA methylation is essential for mammalian embryonic development and aberrant DNA methylation is implicated in several human diseases, in particular in neuro-developmental syndromes (such as the fragile X and Rett syndromes) and cancer [4-6]. These biological significances disclose the importance of understanding genomic patterns and function role of DNA methylation in human, as a initial step to get to know the epigenotype and its manner in connecting the phenotype and genotype.

Two key papers back in 1975 independently suggested that methylation of CpG dinucleotides in vertebrates could be established *de novo* and inherited through somatic cell divisions by protein machineries of DNA methyltransferases that recognizes hemi-methylated CpG palindromes [1, 7]. They also indicated that the methyl group could be recognized by DNA-binding proteins and that DNA methylation directly silences gene expression. After almost four decades, several key points in these foundation papers are proved to be true. Take the mammalian genome for example, there are several findings indicating the epigenetic repression of gene expression by DNA methylation. These include: 1) X-chromosome inactivation, i.e. the inactivation of one of the two X

chromosomes in female somatic cells [8]; 2) Gene imprinting, the allelic-specific silencing occurring at imprinted genes, a group of mono-allelically expressed genes whose pattern are determined by the parental origin of the alleles. These genes are often found playing important roles in development, cellular proliferation and behavior [9, 10]; 3) Suppress the proliferation of transposable elements and repeat elements of viral or retroviral origin [11]. In addition to these, many novel roles of DNA methylation have also been revealed. For example, intragenic methylation has been suggested playing major roles in regulating cell context-specific alternative promoters in gene bodies [12]. DNA methylation can also regulate alternative splicing by preventing CTCF, an evolutionarily conserved zinc-finger protein, binding to DNA [13]. By using the technique of fluorescence resonance energy transfer (FRET) and fluorescence polarization, DNA methylation has also been shown to increase nucleosome compaction through DNA-histone contacts [14]. What is more important, DNA methylation is essential for mammalian embryonic development and aberrant change of DNA methylation has been related to disease such as cancer [15, 16]. However, it is also notable there are several lines of evidence contradicting the relationship between DNA methylation and gene silencing. For example, comparison of DNA methylation levels in human genome on the active and inactive X chromosomes showed reduced methylation specifically over gene bodies on inactive X chromosomes [17, 18]. Not only in human, DNA methylation is found to be usually targeted to the transcription units of actively transcribed genes in invertebrate species [19-21] These results prove that the function of DNA methylation is challenging to be unravel. Besides, due to the development of sequencing technique, whole genome DNA methylation profiles have been detected in

diverse species. Comparing genomic patterns of DNA methylation shows considerable variation among taxa, especially between vertebrates and invertebrates. However, even though extensive studies reveal the patterns and functions of DNA methylation in different species, in the mean time, they also highlight the limits to our understanding of this complex epigenetic system. During my Ph.D., in order to perform in-depth studies of DNA methylation in diverse animals as a way to understand the complexity of DNA methylation and its functions, I dedicated my efforts in investigating and analyzing the DNA methylation profiles in diverse species, ranging from insects to primates, including both model and non-model organisms. This dissertation, which constitutes an important part of my research, mainly focuses on the DNA methylation profile in primates including human and chimpanzee. In general, I will use three chapters to elucidate my work in generating and interpreting the whole genome DNA methylation data. Firstly, we generated nucleotide-resolution whole-genome methylation maps of the prefrontal cortex of multiple humans and chimpanzees, then comprehensive comparative studies for these DNA methylation maps have been performed, by integrating data on gene expression as well. This work demonstrates that differential DNA methylation might be an important molecular mechanism driving gene-expression divergence between human and chimpanzee brains and also potentially contribute to the human-specific traits, such as evolution of disease vulnerabilities. Secondly , we performed global analyses of CpG islands (CGIs) methylation across multiple methylomes of distinctive cellular origins in human. The results from this work show that the human CpG islands can be distinctly classified into different clusters solely based upon the DNA methylation profiles, and these CpG islands clusters reflect their distinctive nature at many biological levels,

including both genomic characteristics and evolutionary features. Moreover, these CpG islands clusters are non-randomly associated with several important biological phenomena and processes such as diseases, aging, and gene imprinting. These new findings shed lights in deciphering the regulatory mechanisms of CpG islands in human health and diseases. At last, by utilizing the DNA methylome from human sperm and genetic map generated from the International HapMap Consortium project, we investigated the hypothesis suggesting a potential role of germ line DNA methylation in affecting meiotic recombination, which is essential for successful meiosis and various evolutionary processes. Even thought the results imply that DNA methylation is a important factor affecting regional recombination rate, the strength of correlation between these two is not as strong as the previous report [22]. Besides, high-throughput analyses indicate that other epigenetic modifications, tri-methylation of histone 3 lysine 4 and histone 3 lysine 27 are also global features at the recombination hotspots, and may interact with methylation to affect the recombination pattern simultaneously. This work suggests epigenetic mechanisms as additional factors affecting recombination, which cannot be fully explained by the DNA sequence itself. In summary, I hope the results from these work can expand our knowledge regarding the common and variable patterns of DNA methylation in different taxa, and shed light about the function role and its major change during animal evolution.

CHAPTER 1

INTRODUCTION

Back to 1958, Francis Crick, who is most noted as the co-discoverer of the structure of the DNA molecule, proposed the famous statement of the central dogma of molecular biology. This explanation of the flow of genetic information within a biological system, which was re-stated in a Nature paper published in 1970 [23], built the bridge for connecting the genotype and phenotype, through the sequential information-carrying biopolymers in living organisms. Since then, major efforts have been dedicated in the identification of genetic mutations, their use as biomarkers, and the understanding of their consequences on human health and well-being. Besides, many comparative studies have been performed between humans and non-human primates at the molecular level to reveal the genetic basis of human specializations [24, 25]. These works suggested that many of the key phenotypic differences among primates mainly result from alterations in the regulation of genes rather than in their sequences [26]. In general, most mechanism studies of the phenotype differences mainly stay at the genomic level by focusing on the regulatory region. For example, DNA-binding transcription factors, which used to be thought as the most crucial determinants of gene expression patterns, can choose genes for transcriptional activation or repression by recognizing the sequence of DNA based in their promoter regions. However, the genotype of transcription factors alone are not sufficient to define the spectrum of gene activity in view of the stable manner of the transcriptional potential of a genome during development [27]. In the past decade, there is an emerging interest in the possibility that changes at levels other than the genetic information could also have long-lasting consequences to the phenotype. These changes usually involve covalent modifications both in DNA and amino acids that constitute the N-terminal tails of histones. These processes are less irrevocable than

genetic mutation and fall under the term 'epigenotype' that can stably maintain patterns of gene expression without changes in DNA sequence. More and more evidence shows that epigenotype constitute a dynamic link between the genotype and the phenotype, both at the stage of establishment in different lineages of the embryo and the stage of somatic maintenance [28]. Modified by many different intrinsic and environmental factors, the resulting epigenotype can determine whether genes are maintained in a repressed or potentially active state, which in turn influences the phenotype both during development and postnatal life. Besides, being heritable and less irrevocable than genetic mutation, epigenotype more likely stands for mark of developmental history since the genomic sequence of a differentiated cell is thought to be identical to the zygote from which it is descended.

As one of the most well-studied epigenotype, DNA methylation is best known as its significance in regulation of gene expression. We hypothesize that changes of DNA methylation may play important roles in regulatory divergence between closely related species. In the first chapter, by generating whole genome, single-CpG resolution DNA methylation profiles in prefrontal cortex of humans and chimpanzees through methyl-C-seq method, we performed a comprehensive comparison of DNA methylation in these closely related species, and provided an unbiased view of the evolution of gene regulation in the context of conservation or changes in DNA methylation profiles.

MATERIALS AND METHODS

Generating Methyl-C-Seq Libraries

Regions of prefrontal cortex were dissected out of postmortem brains of three humans (*Homo sapiens*) and three chimpanzees (*Pan troglodytes*). Chimpanzee samples came from animals that died of natural causes or were euthanized for humane reasons at

the Yerkes National Primate Research Center, and all procedures involving these animals conformed to guidelines established by the Yerkes Institutional Animal Care and Use Committee. Human brain samples were obtained from the Maryland Brain and Tissue Bank from individuals who died of causes unrelated to neurological disorders.

Methyl-C-seq libraries for Illumina sequencing were custom constructed (Alpha Biolaboratory, Burlingame, CA) according to Lister et al. [29] with minor modifications. In brief, ~1 μg of genomic DNA was fragmented by sonication, end repaired, and ligated to custom-synthesized methylated adapters (Eurofins MWG Operon, Huntsville, AL) according to the manufacturer's (Illumina, San Diego, CA) instructions. Adaptor-ligated libraries were subjected to two successive treatments of sodium bisulfite conversion with the EpiTect Bisulfite kit (QIAGEN, Valencia, CA) as outlined in the manufacturer's instructions. Five to ten nanograms of bisulfite-converted libraries was PCR amplified with the following condition: 2.5 U of ExTaq DNA polymerase (Takara), 5 ml of 10XExtaq reaction buffer, 25 mM dNTPs, 1 ml Primer 1.1, and 1 ml Primer 2.1 (50 ml final). The thermo cycle was as follows: 95 ℃ for 3 min and then 14–16 cycles each of 95 ℃ for 30 s, 65 ℃ for 30 s, and 72 ℃ for 60 s. The enriched libraries were purified twice with the solid-phase reversible immobilization method with AMPure beads (Beckman Coulter, Brea, CA). We assessed the library quality by randomly subcloning and sequencing ~20–30 colonies to check for proper library construction and bisulfite conversion. The quality-controlled bisulfite-converted methyl-C-Seq libraries were then sequenced at the UC Berkeley Genome Center and Emory Genome Sequencing Laboratory with the Illumina Genome Analyzer II and the Illumina Hi-Seq, respectively. After quality control, the reads per lane ranged between 15 and 70 million reads. The average phred quality score for each read was 37.

Mapping and Annotation

We first converted all C's to T's both in the reads and in the reference genomes, and we then aligned the converted reads to the converted reference genomes by using the Bowtie algorithm [30]. The assembly versions of the reference genome we used for mapping are GRCh37/HG19 for humans and CGSC2.1/panTro2 for chimpanzees. Total mapped reads accounted for 1.03×10^{11} (humans) and 9.80×10^{10} (chimpanzees) nucleotides, providing 34.33 and 32.63 species-level coverages for human and chimpanzee haploid genomes, respectively.

For comparative analyses of human and chimpanzee methylation profiles, we utilized the data sets from the Chimpanzee Sequencing and Analysis Consortium [31], consisting of 13,454 human-chimpanzee orthologous gene pairs. The orthology of these gene alignments was considered unambiguous and covered the whole coding region. On the basis of these ortholog RefSeq gene IDs, we downloaded the genomic coordinates from the UCSC genome browser. Promoters were defined as regions 1.5 kb upstream and 0.5 kb downstream of the transcription start sites. Gene bodies were defined as those encompassing the region from the transcription start site to the transcription end site. GeneTrail [32] and the DAVID tools [33] were used for the functional annotation enrichment and disease association tests.

Identification of Methylated Cytosines Accounting for False-Positive Rates

We estimated the error rate (nonconversion rate plus sequencing error frequency), p, from the number of cytosine bases sequenced in reference cytosine positions in the unmethylated Lambda genome. Error rates estimated from these were between 0.0013 and 0.0017. We controlled the number of false-positive methylcytosine calls below 0.1% of the total number of methylcytosines as follows: the minimum threshold number of cytosines sequenced at each reference cytosine position at which the position could be called as methylated is equal to $(n * p) / (\alpha(1 - p) + p)$, where n is the read depth for that

site, p is the error rate, and α is a predefined false-discovery value (0.001 for our case). Levels of DNA methylation were calculated by two methods. First, in a false-discovery rate (FDR) method, each reference cytosine was examined and labeled as methylated or unmethylated according to the criterion that the number of false-positive methylcytosine calls should be below 0.1% (see above). In the second method, we calculated the ‘‘fractional methylation’’ values of each cytosine [21, 29]; these values are defined as the total number of ‘‘C’’ reads / (total number of ‘‘C’’ reads + total number of ‘‘T’’ reads). Results from these two methods were highly similar, and the results from the latter method are shown in the main text unless otherwise specified. We discarded those sites with read depths of less than 3. Results from before or after duplicates were removed with the Rmdup tool in the Samtools package were highly similar.

Digital Gene-Expression Profiling Data

Frozen tissue samples from postmortem brains of six humans and six chimpanzees were used. Human and chimpanzee individuals died of causes unrelated to neurological disorders. Samples were dissected either from fresh tissue at the time of brain procurement or later on dry ice from frozen tissue pieces from the frontal pole region of the prefrontal cortex. Total RNA was extracted with QIAGEN’s RNeasy or miRNeasy kits according to the manufacturer’s instructions. All RNA samples were examined for quantity and quality by NanoDrop and Bioanalyzer (Agilent). Sequencing libraries were generated from DpnII-digested poly-A enriched RNA according to the manufacturer’s (Illumina) instructions. BFAST [34] was used for aligning 20 bp reads to both the genome and RefSeq of the respective species. We allowed up to one mismatch with the reference genome in any location within the read. Only reads that aligned to one location in the genome were used for analysis. Alignments to multiple isoforms of a gene were collapsed across gene symbol, and the maximum number of reads for a given isoform was used. A gene was considered ‘‘present’’ if every individual of a species for

a given brain region had at least two reads aligned to the gene. For differential expression analysis, a gene had to be present in at least one of the species being compared. Reads were normalized with quantile normalization.

To examine whether there were any underlying batch effects in our data, we processed all samples from both species together. Analysis of variance [35] of sample traits via univariate linear regression analysis with the first principal component as outcome revealed that species was the most significant sample covariate and was followed by individual and then age. Technical variation sources, including postmortem interval, RNA batch, run batch, and library batch, were not significant, similar to a previous study [36]. Statistical significance of differentially expressed genes was determined with a Bayesian t test. We also performed a two-sample permutation test between human and chimpanzee expression values and compared it to the p value from our original method. At the 5% significance level, approximately 92% of genes showed a concordant pattern between these two methods. For the inconsistent genes, most were significant from the permutation test and weakly significant from our original method.

Comparative Human Methylome Analysis among Different Tissues

We compared the human prefrontal cortex (brain) methylome that had the highest mean read depth and lowest duplicate read count (Hs1570) to methylomes generated from human embryonic stem cells (ESCs) [37], human neonatal foreskin fibroblasts [37] and human peripheral-blood mononuclear cells (PBMCs) [38]. Methylation data from other tissues and cell lines were obtained from respective publications. In brief, the ESCs were derived from aWA09 hESC line, and were cultured feeder free on Matrigel (Becton Dickinson) in StemPro medium (Lifetech), and were passaged with Accutase (Lifetech). The neonatal fibroblast cell lines were obtained from GlobalStem (newborn human foreskin fibroblasts, untreated) and were harvested for analysis at passage. The human PBMCs were obtained from the same individual as in the YanHuang project, which is the

first finished diploid genome sequence of an Asian individual. Methylome data on ESCs, neonatal fibroblasts, and PBMCs were downloaded from the Gene Expression Omnibus, and coordinates were converted from human genome build hg18 to hg19 with the UCSC liftover tool. Bisulfite-converted sequence data were merged for all CG dinucleotides and CH dinucleotides (H = A, C, or T) that had at least three strand-specific reads in each of the four methylomes being compared. Mean fractional methylation of annotated elements was calculated as the mean of fractional methylation values for each site within the annotated element.

RESULTS

Genome-wide DNA-Methylation Patterns Reveal Extremely Heavily Methylated Brains

By sequencing the bisulfite-converted genomic DNA from prefrontal cortex samples, we generated whole genome, nucleotide-resolution DNA methylation maps (methylomes) from three humans and three chimpanzees. Sequencing reads representing 1.03×10^{11} and 9.80×10^{10} base pairs were generated from human and chimpanzee prefrontal cortex samples, respectively, corresponding to an average read depth of 11.4X and 10.9X per haploid genome. Bisulfite conversion rates estimated from unmethylated lambda DNA controls show that the conversion rates are high enough to make sure our method faithfully captures patterns of genomic DNA methylation in these samples (Methods). Prefrontal cortex methylation maps from both species revealed extremely heavy CpG methylation, where between 79.4% to 82.5% of CpGs are methylated. In comparison, only minor fractions of non-CpG sites (1.3% to 2.2%) are methylated (Figure 1.1). Among the genomic regions, promoters and CpG islands are generally hypomethylated. Transposable elements are the most heavily methylated in both species (Figure 1.1C), supporting the idea that DNA methylation suppresses proliferation of

transposons in these genomes [3]. DNA methylation levels across transcription units exhibit distinctive patterns similar to previous findings, where DNA methylation levels dip at the transcription start site, increase along the transcribed unit (gene body), and decrease again at the transcription termination site.

To gauge tissue-specific differences in levels of DNA methylation, we compared the methylation maps of the human prefrontal cortex to those from three other tissues, including ESCs, fibroblasts, and PBMCs. These methylomes were all generated using similar methods, facilitating a direct comparison of overall levels of DNA methylation among these tissues [37, 38]. Our analysis reveals that the prefrontal cortex is the most heavily methylated of these four tissues (Figure 1.1A). A high level of methylation of prefrontal cortex is consistent throughout different genomic regions and across different cytosine classes (Figure 1.1).

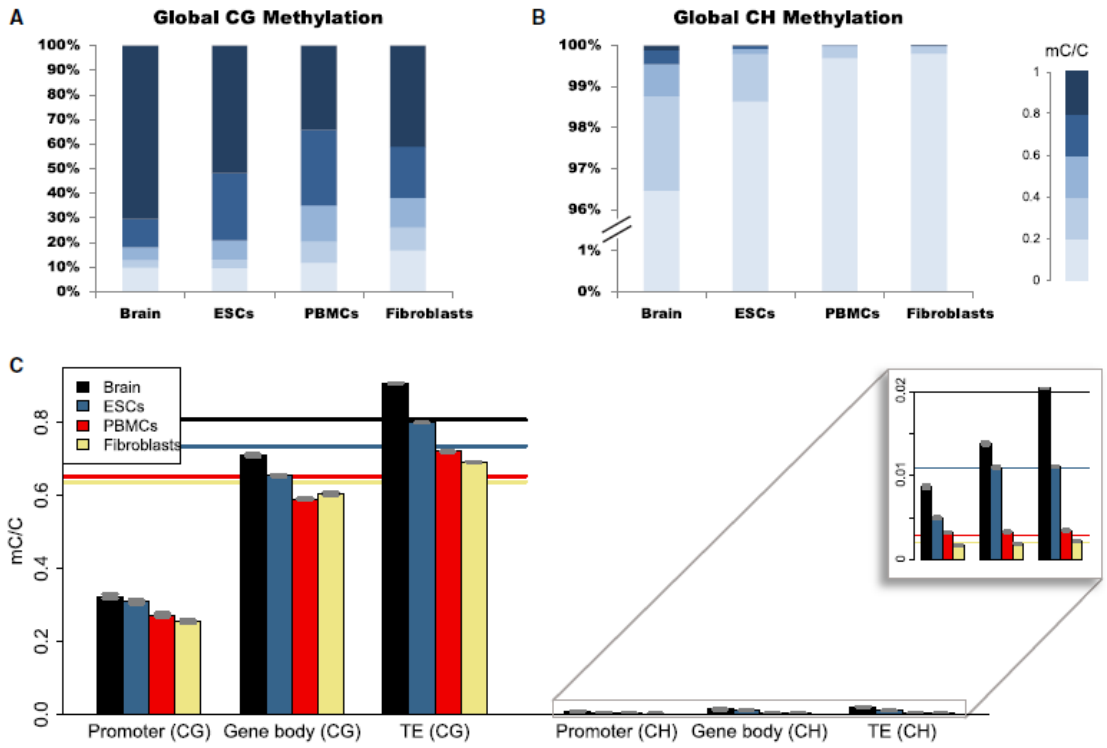

Figure 1.1. Differences in DNA-Methylation Levels among Human Tissues and Genomic Features (A) Proportional representation of genome-wide DNA-methylation levels for individual CG dinucleotides in the human prefrontal cortex (brain), ESCs, neonatal fibroblasts, and PBMCs.(B) Same analyses as in (A) but for CH dinucleotide context (H = A, T, or C). (C) Mean methylation levels in each tissue for gene promoters (CG context, n = 18,416; CH context, n ? 18,584), gene bodies (CG context, n = 18,477; CH context, n = 18,656), and transposable elements (CG context, n = 1,837,431; CH context, n = 2,989,765). Horizontal lines indicate global means of methylation levels for individual CG sites (main panel) or CH sites (inset). Error bars indicate 95% confidence intervals of the mean.

Interspecies and Intraspecies Variation of Genome-wide Patterns of DNA Methylation

Genome-wide brain methylation maps of humans and chimpanzees exhibit intriguing intraspecific and interspecific variation (Figure 1.2). Interestingly, prefrontal cortex samples from younger individuals in our study exhibit higher levels of DNA methylation in both species (Figure 1.2C and 1.2D). For example, the chimpanzee

individuals are 24, 27, and 43 years of age. At the genome-wide level, the third (43-year-old) individual exhibits slightly but significantly lower methylation than the other individuals. In human samples, a younger (31-year-old) individual is overall more heavily methylated than the other two individuals of ages 47 and 48 years. However, given the small sample size, these results should be taken with caution and need to be validated in a study with a larger number of individuals spanning greater variation of ages. In term of the interspecies variations, the degree of DNA methylation is also slightly but significantly different between human and chimpanzee brains. At the whole genome level, the average fractional methylation levels of CpG dinucleotides in the human and chimpanzee genomes are 80.9% (±0.036%) and 82.1% (±0.034%), respectively (Mann-Whitney test, $P < 10^{-15}$). In addition, species differences in DNA methylation levels are also apparent in both promoters and gene bodies of 12,533 human-chimpanzee orthologs by using principal-component analyses (Figure 1.2A and 1.2B). Thus, our data suggest that human prefrontal cortex regions are generally less methylated than chimpanzee prefrontal cortex regions. Our findings are at odds with a previous study that reported the opposite trend based upon a limited number of CpG sites [39]. However, an analysis of the specific CpG sites included previously revealed no difference between the two species in our data. The difference might in part be due to the fact that the previous study used a low-resolution methylation array developed specifically for the human genome. Moreover, our genome-wide results are consistent with another earlier study using HPLC, which suggested that human brains are generally less methylated than brains of other primates [40].

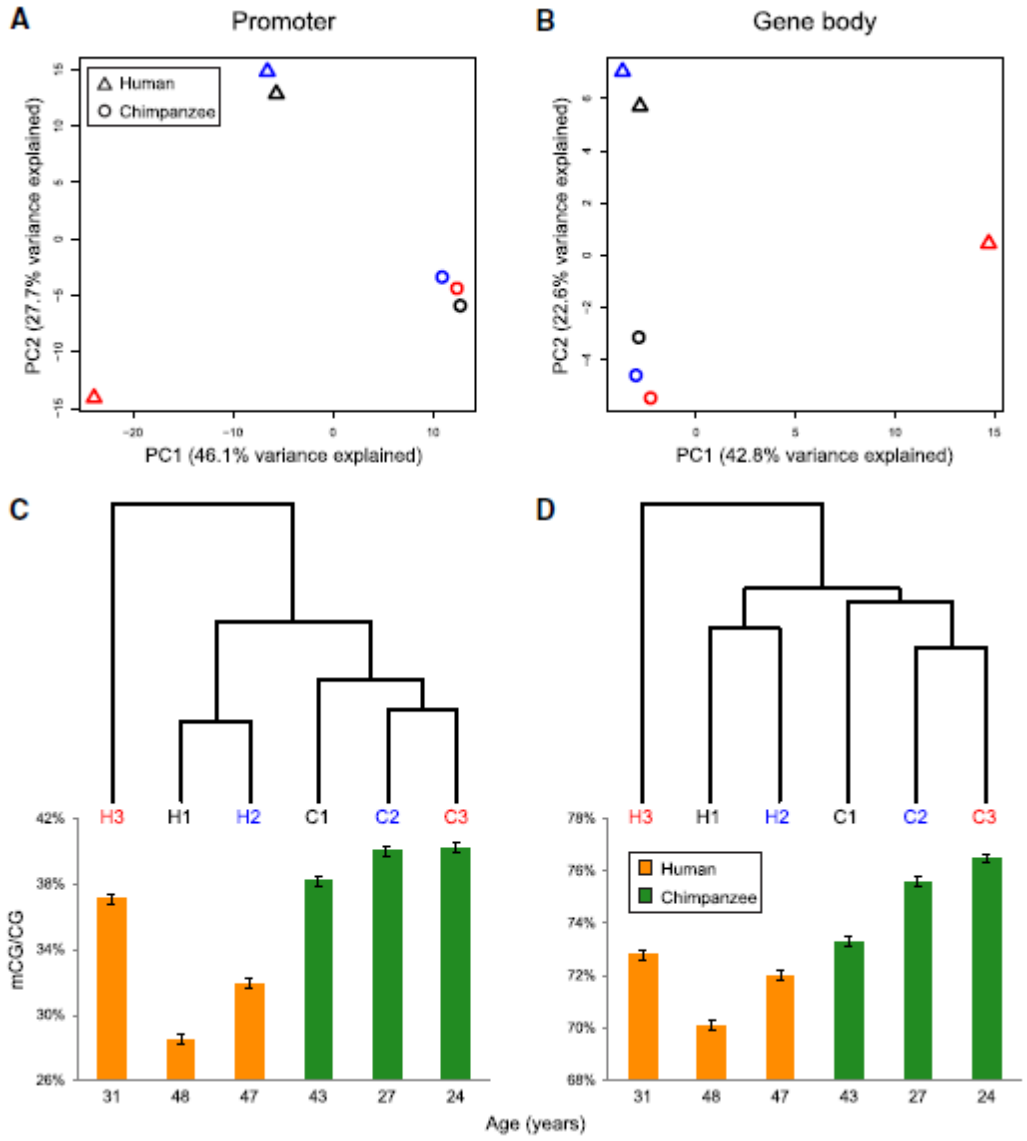

Figure 1.2. Between- and Within-Species Variation of Genomic DNA Methylation in Human and Chimpanzee Prefrontal Cortex Regions. Principal-component analyses of (A) promoters and (B) gene bodies of human-chimpanzee orthologs demonstrate that the patterns of DNA methylation are distinct between humans and chimpanzees. For promoters, the first principal component, which explains 46.1% of variation, distinguishes samples from human and chimpanzees. The second principal component, explaining 27.7% of total variation, separates two human samples from the third one. For gene bodies, the first principal component (explaining 42.8% of total variation) separates the third human from the rest, whereas the second principal component (explaining 22.6% of total variation) separates the human and chimpanzee brains. Hierarchical clustering analyses of (C) promoters and (D) gene bodies demonstrate that the overall levels of methylation are lower in human brains than in the chimpanzee brains. The youngest human individual (H3) exhibits the most distinctive

pattern of DNA methylation. The error bars indicate 95% confidence intervals of the mean.

Distinctive Patterns of Promoter Methylation, Functional Enrichment, and Disease Association

Previous studies determined that DNA methylation in vertebrate promoters occurs in a discrete fashion—these promoters can be classified as hypermethylated and hypomethylated [41, 42]. In accordance with these studies, promoter DNA methylation in human and chimpanzee brains falls into distinct hypermethylated and hypomethylated classes (Figure 1.3A). In comparison, gene bodies are generally heavily methylated in the prefrontal cortex of both species (Figure 1.3B), which is expected under ‘‘global’’ patterns of genomic DNA methylation [41, 43]. Levels of DNA methylation in promoters and gene bodies are clearly lower in the human brain than in the chimpanzee brain(Figure 1.3A and 1.3B), a difference that is especially marked for promoters (Figures 1.3A and 1.3C), which on average exhibit 23% less methylation in humans than in chimpanzees.

To identify significantly differentially methylated promoters between human and chimpanzee brains, we performed the following tests. First, we performed a Fisher's exact test by using the total numbers of methylated and unmethylated CpG sites in all samples and calculated adjusted p values by the FDR method for multiple testing [44, 45]. Then, from the pool of significantly differentially methylated promoters obtained by this test, we further classified genes into those with hypermethylated (defined as fractional methylation levels > 0.8) or hypomethylated (fractional methylation levels < 0.2) promoters (Figure 1.3). From these gene sets, we identified 474 genes whose promoters had ‘‘switched’’ between the hypermethylated and hypomethylated classes between the human and chimpanzee brains. In the majority (n = 468) of these promoters, human brains exhibit conspicuously lower levels of DNA methylation than do chimpanzee brains. Interestingly, these genes are significantly enriched in molecular

functions such as protein binding and phosphotransferase activity (Table 1.1). Moreover, they exhibit striking associations with several disorders, including neurological and psychological disorders and cancers. For example, genes whose variants are associated with autism are 3.5-fold enriched in this group of genes (although not significantly so because of the small number of genes (Table 1.2)).

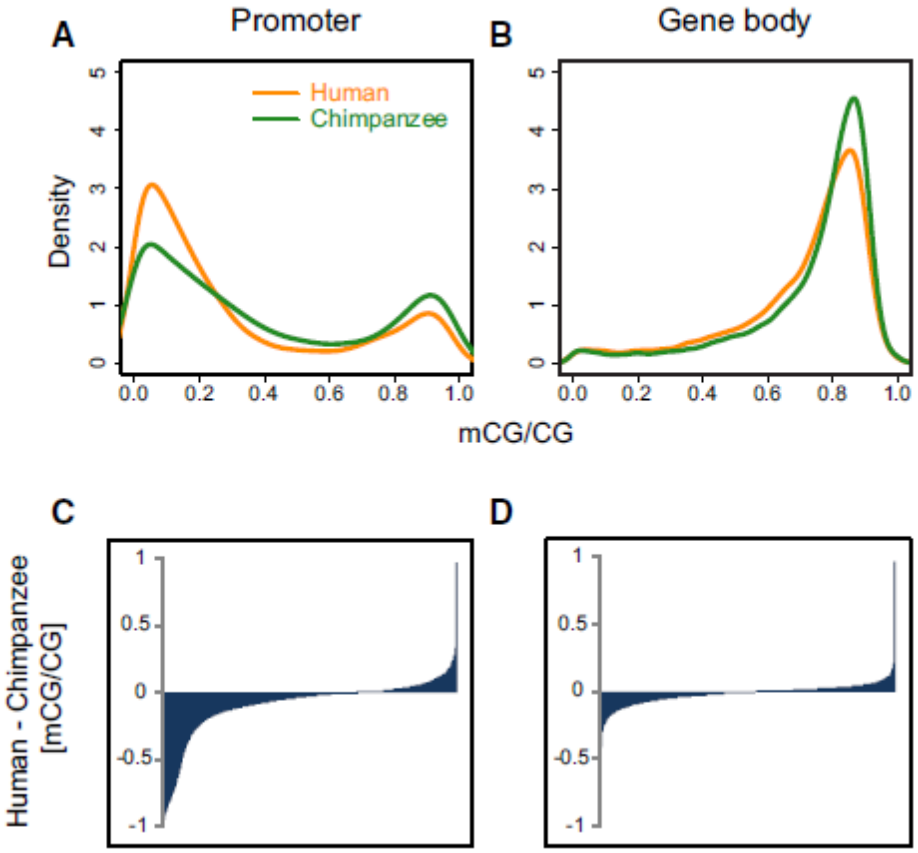

Figure 1.3. Patterns of DNA Methylation in Genic Regions Influence Gene Expression. Density plots of (A) promoter and (B) gene-body DNA methylation from humans and chimpanzees. Promoter DNA methylation exhibits distinctive ‘‘bimodal’’ patterns. In comparison, gene bodies of both species are heavily methylated (B). DNA-methylation-level differences, measured as the mean of human methylation levels minus the mean of chimpanzee methylation levels, show that promoters particularly exhibit lower levels of DNA methylation in the human brain than in the chimpanzee brain (C). In contrast, gene bodies show similar levels of DNA methylation between species (D).

The above-described method for identifying differentially methylated promoters is perhaps overly stringent. Thus, we developed a second method, based on the relative

difference in promoter methylation, to identify differentially methylated promoters. Beginning with genes for which Fisher's exact test using the false discovery method was significant, we first defined genes whose relative methylation levels have changed more than 50% (in other words, |(Chimp fractional methylation level - Human fractional methylation level)/(Chimp fractional methylation level + Human fractional methylation level)| is greater than 0.5). We further restricted analysis to genes for which the absolute difference between the fractional methylation levels of humans and chimpanzees is greater than 0.2. Using this method, we identified 1055 genes that are significantly less methylated in the human brains compared to the chimpanzee brains. Analyses of these promoters again demonstrate patterns of functional enrichment and disease association similar to the above results.

Table 1.1. Genes whose promoters are hypo-methylated in the human brains while hyper-methylated in the chimpanzee brains (n=468) are enriched in specific gene ontology (GO) terms.

GO terms	Accession	P-value (FDR)
cellular process	GO:0009987	7.2e-05
protein binding	GO:0005515	1.8e-04
cellular macromolecule metabolic process	GO:0044260	1.9e-03
cellular metabolic process	GO:0044237	4.3e-03
transferase activity, transferring phosphorus-containing groups	GO:0016772	1.0e-02

Table 1.2. Disease genes found in genes whose promoters are hypo-methylated in the human brains while hyper-methylated in the chimpanzee brains (n=468).

Category	Count	Fold Enrichment	Genes
neural tube defects	5	4.7	*PDGFRA*(MIM 173490), *SHMT1*(MIM 182144), *TYMS*(MIM 188350), *DHFR*(MIM 126060), *CXCL6*(MIM 138965)
Autism	6	3.5	*GABRA2*(MIM 137140) , *GSTM1*(MIM 138350) , *SLC6A4*(MIM 182138), *ACCN1*(MIM 601784), *CLOCK*(MIM 601851), *GABRG1*(MIM 137166)
alcohol dependence	4	5.0	*GABRA2*(MIM 137140), *SLC6A4*, *GABRB1*(MIM 137190), *GABRG1*
Chemodependency	9	2.0	*GABRA2*, *GSTM1*, *SLC6A4*, *GABRB1*, *CLOCK*, *SCN5A*(MIM 600163), *HOMER1*(MIM 604798), *GABRG1*, *CRTC1*(MIM 607536)
Cancer	27	1.3	*HPSE*(MIM 604724), *IRAK4*(MIM 606883), *TES*(MIM 606085), *KIT*(MIM 164920), *RECQL*(MIM 600537), DHFR, *KDR*(MIM 191306), *IKZF3*(MIM 606221), *RAD51D*(MIM 602954), *CDK4*(MIM 123829), *CSF1*(MIM 120420), *LIG3*(MIM 600940), *SUOX*(MIM 606887), *CXCL5*(MIM 600324), *NRAS*(MIM 164790), *PDGFRA*, GHR, *RASSF8*(MIM 608231), *TYMS*(MIM 188350), *POLR2B*(MIM 180661), *VDR*(MIM 601769), *SLC6A4*, *GSTM1*, *SHMT1*(MIM 182144), *STARD3*(MIM 607048), *IGFBP7*(MIM 602867), *POLK*(MIM 605650)

DNA Methylation and Gene Expression in the Human and Chimpanzee Brains

A well-known consequence of DNA methylation is its effect on the regulation of gene expression [46]. Furthermore, differential expression of genes in humans and chimpanzees may drive lineage-specific patterns of evolution [25, 26, 47]. Given the profound influence of promoter DNA methylation on the regulation of gene expression, we asked whether changes of DNA methylation might underlie gene expression divergence between human and chimpanzee brains. To address this question we integrated data on DNA methylation with data on gene expression from human and chimpanzee prefrontal cortex, generated using a next-generation sequencing method, digital gene expression profiling (DGEP, see Materials and Methods).

Levels of DNA methylation from promoters and gene bodies are each significantly negatively correlated with levels of gene expression (Spearman's correlation coefficients range between -0.18 ~-0.24, as shown in Figure 1.4). Several recent studies demonstrated a 'bell shape' relationship between gene expression and methylation, where the most heavily methylated gene bodies are often expressed at intermediate levels, and genes expressed at high and low levels are moderately methylated [20, 21]. However, in the prefrontal cortex samples, gene body methylation decreases roughly linearly with increasing levels of gene expression in both species (Figures 1.4B and 1.4D). This finding is similar to a recent study where a linear and negative relationship between gene expression and DNA methylation in brain (both the occipital lobe and whole brain) was reported [46]. Thus the effect of gene body DNA methylation and gene expression is not universal across different tissues.

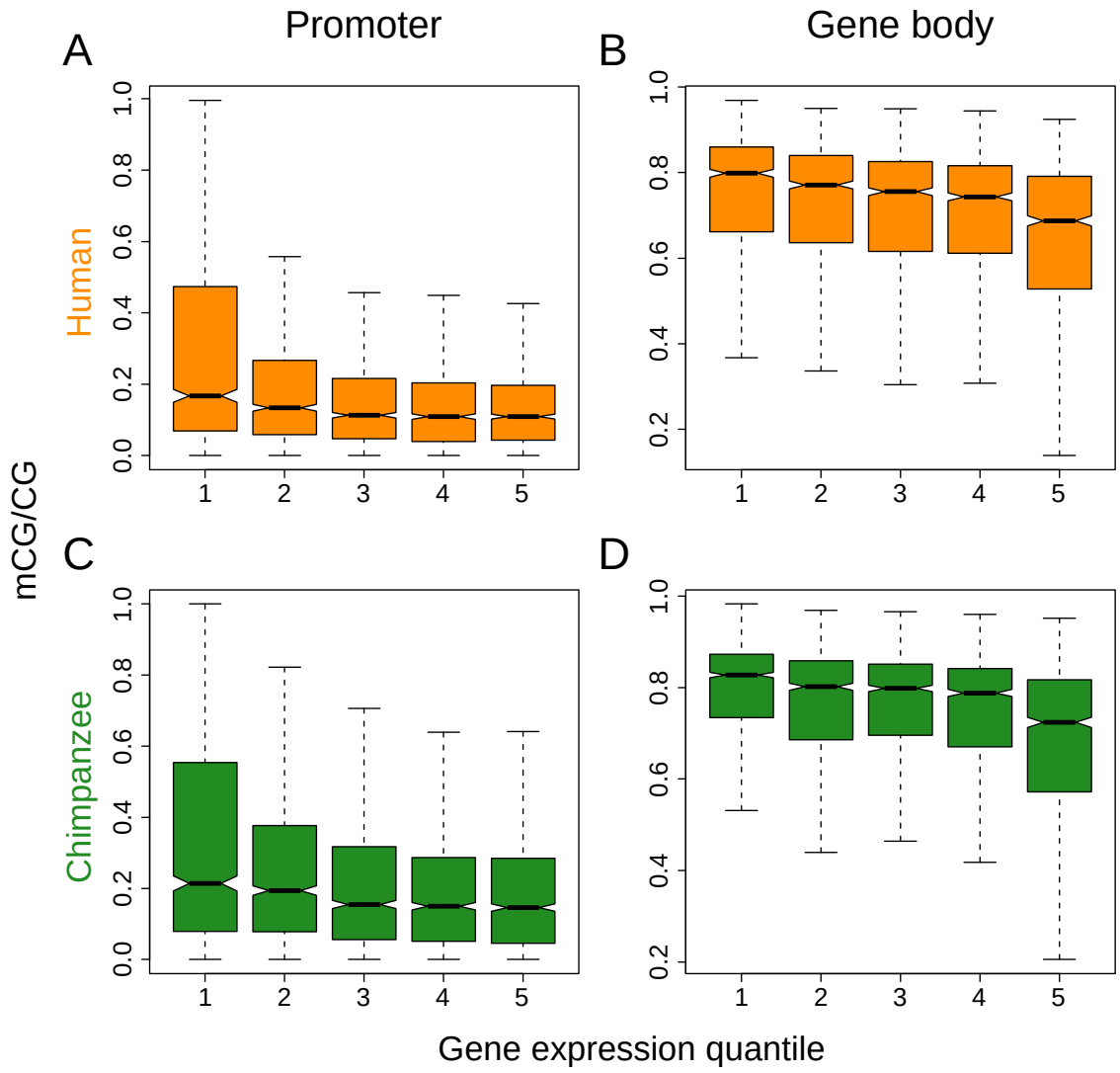

Figure 1.4. DNA methylation is negatively correlated with gene expression level in both promoters and gene bodies in prefrontal cortex. Integrating levels of DNA methylation with levels of gene expression measured by digital gene expression profiling, we observe a negative correlation between human gene expression level and (A) human promoter methylation (Spearman's correlation coefficient r = -0.24, $P < 10^{-15}$) as well as (B) human gene body methylation (r = -0.18, $P < 10^{-15}$). The X-axis represents increasing levels of gene expression from left to right. We also observe a negative correlation between chimpanzee gene expression level and (C) chimpanzee promoter methylation (r = -0.19, $P < 10^{-15}$) as well as (D) chimpanzee gene body methylation (r = -0.20, $P < 10^{-15}$).

Among the genes whose promoters are hypo-methylated in human but hyper-methylated in chimpanzee brains, expression-level data are available for 273 genes. A majority of these exhibit higher expression in human brains compared to chimpanzee

brains (168 out of 273, $P < 10^{-4}$, binomial test). In comparison, none of the three genes whose promoters are hyper-methylated in humans compared to chimpanzees exhibit increased expression in humans. When we restrict our analyses to genes with expression patterns that are significantly different between human and chimpanzee brains (Bayesian t-test, $P < 0.05$), the same pattern is observed: 41 out of 58 genes with significantly hypo-methylated promoters in humans compared to chimpanzees exhibit higher levels of expression in human ($P < 10^{-4}$). Thus, differential promoter methylation between humans and chimpanzees manifest in different transcriptional levels (Figure 1.5A).

Again we find that many of these genes are implicated in neurological functions and disorders (Figure 1.5B). For example, the insulin-like growth factor binding protein 7 (*IGFBP7* [MIM 602867]) gene regulates insulin-like growth factor availability and receptor binding, and is implicated in extinction of fear memories and neuro-genesis [48]. Methylation levels of IGFBP7 promoters are dramatically different between the human and chimpanzee brains, and the expression of this gene exhibits a pattern concordant with the methylation pattern (Figure 1.5). In another example, the sodium channel, voltage gated, type VIII alpha subunit (*SCN8A* [MIM 600702]) gene is implicated in wide-ranging neurological and behavioral disorders and cognitive impairment, [49, 50] and is also hypo-methylated and significantly more strongly expressed in the human brain compared to chimpanzee brain (Figure 1.5).

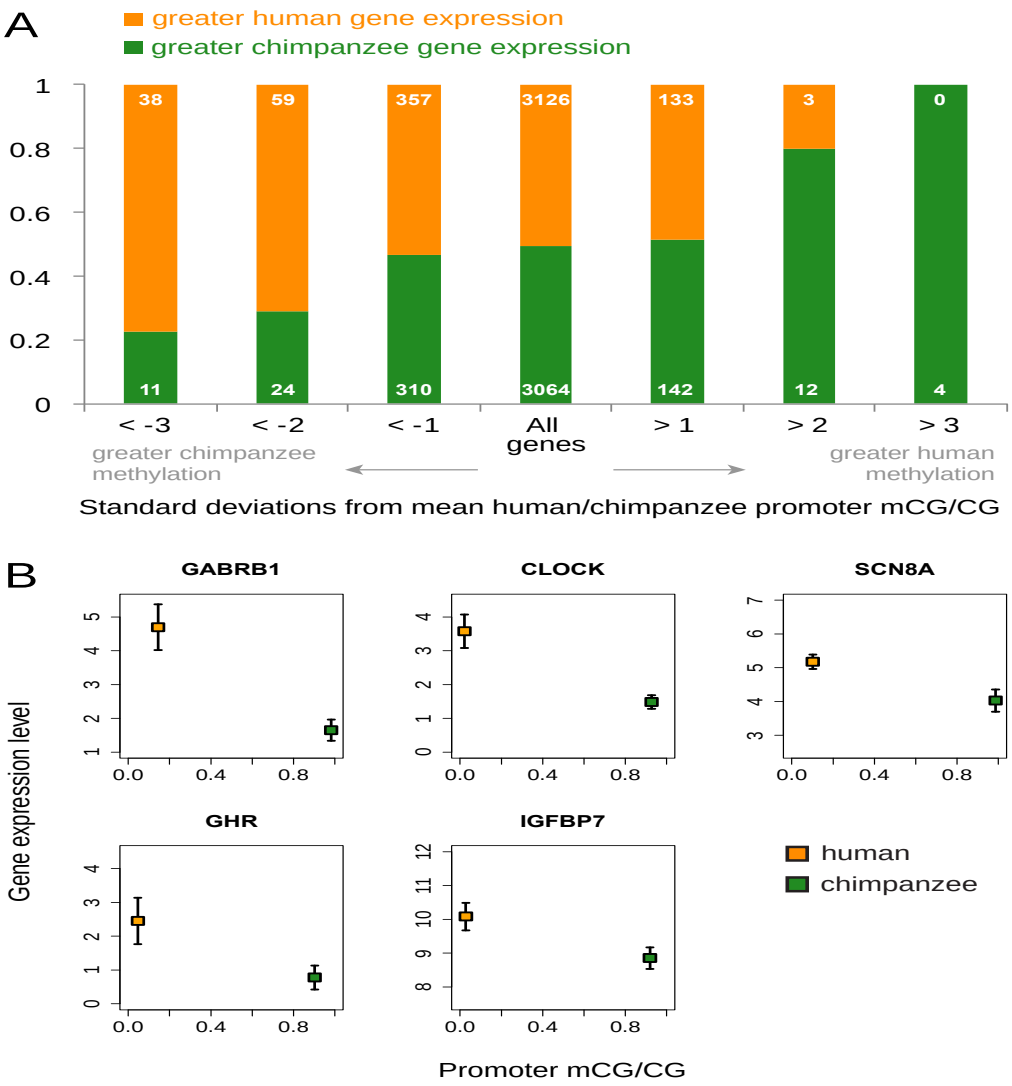

Figure 1.5. Differences in promoter methylation associated with differences in gene expression between human and chimpanzee prefrontal cortex. (A) The proportion of genes with higher or lower expression values in human, as compared to chimpanzee, prefrontal cortex. Each bar represents a class of genes based on the number of standard deviations from the mean ratio of methylation measures in human versus chimpanzee promoters in the prefrontal cortex. (B) Selected genes with hypo-methylated promoters in human, hyper-methylated promoters in chimpanzee, and significantly higher expression in human than chimpanzee. Error bars, 95% confidence intervals of the mean (n = 6). Gamma-aminobutyric acid (GABA) A receptor, beta 1 (GABRB1) is involved in neurotransmission of the central nervous system. Clock homolog (mouse) (CLOCK) encodes a transcription factor essential to the circadian rhythm. Sodium channel, voltage gated, type VIII, alpha subunit (SCN8A) facilitates the generation of action potentials in neurons and other cells. Growth hormone receptor (GHR) is integral to activating insulin-like growth factor production, leading to growth. Insulin-like growth factor binding protein 7 (IGFBP7) regulates insulin-like growth factor availability and receptor binding.

DISCUSSION

Recent technical advances have enabled us to examine genomic variation of DNA methylation at the nucleotide-level [21, 29], revealing highly complex and dynamic tissue- and cell type- specific patterns of genomic DNA methylation. In parallel, new functional studies are illuminating multi-faceted connections between DNA methylation and regulation of gene expression. In addition to the well-known effect of promoter methylation in silencing gene expression, DNA methylation is also implicated in the regulation of alternative splicing [12] and the regulation of miRNA [51]. Thus, DNA methylation harbors a strong potential to influence regulatory divergence between species.

To elucidate the evolutionary significance of DNA methylation, in this study we examined the differences in genome-wide DNA methylation maps of human and chimpanzee brains and their consequences on gene expression divergence. A few studies have previously investigated methylation difference between humans and non-human primates, but these studies either examined an extremely limited number of sites or used methods that are low-resolution and potentially biased due to underlying sequence differences [39, 40, 52]. In contrast, our study used the methyl-C seq method to resolve detailed patterns of genomic DNA methylation at individual nucleotide resolution.

One of the advantages of the methyl-C seq method is that it allows us to infer methylation frequencies of individual CpGs quantitatively. Our DNA methylation maps reveal the prefrontal cortex to be by far the most heavily methylated tissue investigated so far (Figure 1.1). Our results stand in contrast to the hypothesis that DNA methylation decreases in conjunction with cellular differentiation [37]. Rather, our study suggests that DNA methylation patterns undergo dynamic reprogramming in a tissue and cell-type specific manner. The striking enrichment of DNA methylation in brains (Figure 1.1) also has important evolutionary implications. It has been shown repeatedly that genes expressed in brains are, on average, the most evolutionarily constrained both in terms of

sequence evolution as well as gene expression patterns [25, 53]. The observation that the brain is the most heavily methylated among the tissues investigated so far suggests that DNA methylation may contribute to the constraints on sequence and expression evolution, possibly by suppressing gene expression noise [54]. Similarly, heavy methylation of transposable elements in brain may indicate particularly strong silencing of transposable elements [3].

We observed intriguing within- and between-species variation of DNA methylation in the brains of humans and chimpanzees. In both species, samples from younger individuals (31 versus 47 and 48 in humans, 24 and 27 versus 43 in chimpanzees) tend to exhibit heavier DNA methylation compared to older individuals (Figure 1.2). Previous studies, investigating limited numbers of CpG sites or genes, reported both increases and decreases of DNA methylation with aging [55-59]. Our data, while representing the first genome-wide analyses of CpG sites, consist of only three individuals per species with relatively similar ages, thus should be taken with caution. Nevertheless, it is interesting to note that studies analyzing CpG islands have generally reported increased DNA methylation with increasing age, while some studies reported that CpGs that are not in CpG island context tend to lose DNA methylation with aging [58].

The overall patterns of DNA methylation differ between human and chimpanzee brains: notably the chimpanzee brains exhibited higher DNA methylation levels compared to human brains. Our results are in accord with an earlier study that used high performance liquid chromatography (HPLC) to quantify the levels of methylcytosines from the brains of human, macaque, African green monkey and squirrel monkey and showed that the human brain exhibited the least amount of methylcytosines among these species [40]. However, the fact that DNA methylation varies with age, and that it is not straightforward to 'match' ages between human and chimpanzee samples, cautions drawing a general conclusion from the limited number of samples used in this study.

Nevertheless, it is notable that the species-level difference between humans and chimpanzees is the most pronounced in promoters (Figure 1.3). Given the observation that human promoters are generally hypo-methylated compared to chimpanzee promoters, the increase of gene expression in the human brains compared to the brains of chimpanzees [60-62] may be partially mediated by an overall decrease of DNA methylation, particularly in promoters. Future analyses of outgroup primates, such as Old World monkeys, will help elucidate lineage-specific changes in these epigenetic modifications.

Furthermore, promoters that are significantly differentially methylated between the brains of humans and chimpanzees (most of which are hypo-methylated in the human brains compared to the chimpanzee brains) are enriched in several functional categories, including protein binding and cellular metabolic processes. Strikingly, the list of genes harboring differentially methylated promoters includes disproportionately high numbers of those associated with human diseases (Table 1.2). In particular, this list of disease includes neuro-developmental and psychological disorders, such as neural tube defects, autism, and alcohol and other chemical dependencies. Interestingly, they represent a characteristic set of diseases to which modern humans are particularly susceptible [63]. This suggests that methylation differences between human and chimpanzee brains may have significant functional consequences and potentially bear relevance to the evolution of human specific disease vulnerabilities. Given that DNA methylation functions as a modulator of environmental signals to cellular regulatory machineries [64, 65], comparative epigenomic studies like ours will allow us to better understand both the genetic and environmental contributions to species differences. Thus, our results highlight the utility of comparative studies in identifying key epigenomic modifications underlying human specific phenotypes, including disease vulnerabilities.

CHAPTER 2

INTRODUCTION

Thirty-eight years ago, two independent studies proposed that cytosine DNA methylation in eukaryotes could act as a stably inherited modification affecting gene regulation and cellular differentiation [1, 7]. Since then, intense effort has expanded our understanding of diverse aspects of DNA methylation in higher eukaryotic organisms, especially human. Now it is well known that the degree of DNA methylation in the human genome is extensive: most of CpG dinucleotides in the human genome are methylated in most tissues and developmental stages, which is referred to as "global DNA methylation" [66]. However, some genomic regions, e.g., CpG islands, exhibit prominent exceptions to this pattern [67]. Originally, CpG islands were defined as clusters of hypo-methylated CpG dinucleotides in the heavily methylated mammalian genomes [68, 69]. In order to get a comprehensive map of CpG islands among genomic sequences, several computational algorithms have been developed [70, 71]. A key feature of these computational algorithms is a metric to quantify the observed frequency of CpG dinucleotides normalized by the G+C content, commonly referred to as 'CpG O/E'. Genomic regions exhibiting particularly CpG O/E, among other characteristics, are generally considered as good candidates of CpG islands.

Accompanied with the improvement of CpG islands definition, numerous studies also indicated the critical importance of CpG islands in regulatory and developmental processes. For example, many CpG islands co-localize with promoters [72-74]. They are often characterized by transcriptionally permissive chromatin states [75, 76], and frequently overlap with enhancers and other regulatory elements [77-80].

The regulatory effects of CpG islands often critically rest on the 'correct' (or the lack of) DNA methylation. For example, even though CpG islands are generally

characterized by their unmethylated status, some CpG islands undergo DNA methylation, often in tissue- or developmental stage-specific manner. What is more important, aberrant methylation at some CpG islands is implicated with diseases, in particular, cancer [81, 82]. Therefore, understanding the full extent of variation of DNA methylation in CpG islands and its functional consequences has tremendous implications for advancing our knowledge of molecular mechanisms of regulation and development.

Moreover, recent studies begin to unfold intriguing functional heterogeneity among CpG islands. For example, long CpG islands and short CpG islands exhibit different regulatory activities such as gene expression complexity [83] as well as nucleosome depletion patterns [84]. A recent evolutionary study has determined that while the majority of CpG islands may actively avoid DNA methylation, some CpG islands are likely to maintain high CpG contents via methylation-independent processes such as biased gene conversion [85]. These findings begin to shed lights on the potential diversity among CpG islands. At the same time, they highlight many unanswered, critical questions: for example, do the computationally predicted lists of CpG islands capture the true epigenomic and functional complexity of CpG islands? Do all CpG islands exhibit tissue and developmental stage specific variation of DNA methylation? Alternatively, is there a group of CpG islands that tend to exhibit variable patterns of DNA methylation? How are these variations of DNA methylation related to regulatory functions of CpG islands? Do methylation profiles of CpG islands differ according to their evolutionary mechanisms? In chapter 2, I will describe my investigation of these pressing questions by analyzing whole-genome, nucleotide resolution methylation maps from multiple methylomes of distinctive origins.

MATERIALS AND METHODS

Whole Genome Methylomes

We used whole genome, nucleotide-resolution DNA methylation maps (methylomes). We focused on analyzing normal tissues or primarily tissue derived cell lines, rather than differentiated cell lines or cancer genomes. Primary data consists of methylomes generated from human embryonic stem cells (ESCs) [86], human neonatal foreskin fibroblasts [86], human peripheral-blood mononuclear cells (PBMCs) [87], prefrontal cortex of human brain [88] and human sperms [89]. These methylomes were all generated with next-generation bisulfite sequencing technology and have similar number of mapped CpG sites, facilitating a direct comparison of CpG island methylation among tissues. As a comparison, we contrasted the whole genome methylation data from the prefrontal cortex to those generated via the reduced representation bisulfite sequencing methods as a part of the ENCODE project from the " BC_Brain_H11058N " cell line. Comparison of these data sets demonstrates that the whole genome methylation sequencing provides a superior coverage of CpG islands (Figure 2.1). We extended our analyses to three additional methylomes: placenta, kidney and cerebellum [90]. These three methylomes were generated using the same methods, but are of lower coverages (Table 2.1).

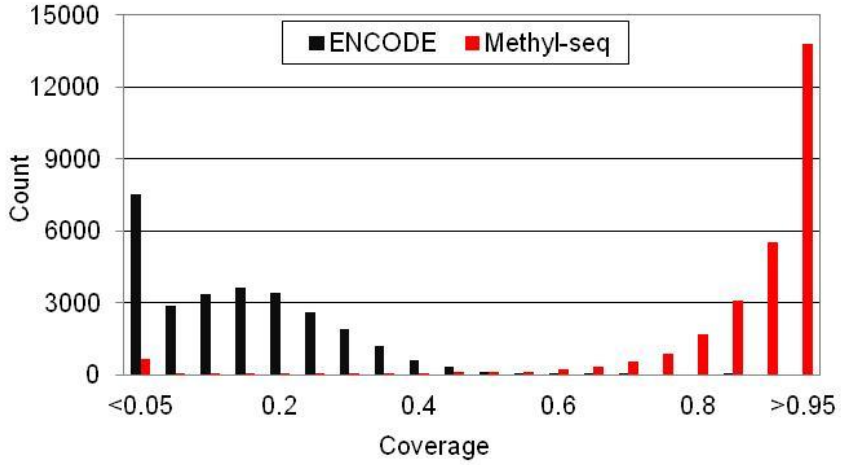

Figure 2.1. CpG island coverages. We compared the coverages of CpG sites in CpG islands between the data generated by the ENCODE project to the data from comprehensive methyl C seq method. The ENCODE project used a modified version of Reduced Representation Bisulfite Sequencing. These profiles present an excellent overview of genomic DNA methylation variation. However, to obtain a comprehensive variation of DNA methylation of CpG islands, the methylC seq data clearly outperforms this data, as shown below. We compared the brain methyl seq data to ENCODE data from the BC_Brain_H11058N cell line. The distribution of CpG sites coverage within CpG islands is shown for the ENCODE data (black) and for the methylC-seq data (red). Coverage is calculated as number of mapped CpGs divided by number of total CpGs in each CpG island.

Table 2.1. Human DNA methylome datasets used in chapter 2

Tissue	Gender	Dataset ID	Coverage	Total Mapped CpGs (millions)	Mapping program[1] and efficiency[2]
Embryonic stem cells	F	GSE19418	9.1X	26.2	SOAP2 (98.5%)
Neonatal foreskin fibroblasts	M	GSE19418	9.8x	26.4	SOAP2 (99.3%)
Prefrontal cortex of brain	M	GSE37202	11.4X	26.8	BS-seeker (99.8%)
Peripheral Blood Mononuclear Cells	M	GSE17972	10X	27.1	SOAPaligner (99.1%)
Sperm	M	GSE30340	16X	28.2	RMAPBS (98.7%)
Placenta	F	GSE39775	1.6X	23.9	BS-seeker (99.3%)
Cerebellum	M	GSE39775	0.3X	8.3	BS-seeker (99.3%)
Kidney	M	GSE39775	0.5X	8.9	BS-seeker (99.3%)

1: The mapping procedures for all tissues used a "reduced"' three-letter alphabet comprising A, G, and T, where all C's were converted to T's both in the reads and in the reference genome to accommodate the conversion of unmethylated cytosines to thymines by bisulfite conversion. Besides, all the datasets contain the quality control steps such as the standard Illumina sequencing pipeline for base calling and quality filtering, all the redundant reads were removed before the alignment to the reference genome.
2: The efficiency is referred as the conversion rate of the unmethylated CpGs from the bisulfite treatment.

CpG Island Annotation and Methylation

The annotations of the CpG island used in this study were downloaded from UCSC Genome Browser [91]. These CpG islands are characterized as at least 200 bps in length, GC content of 50% or greater, and a CpG frequency (observed/expected; [o/e]) of 0.6 and exclusive of repetitive sequences. To estimate the methylation level for each CpG island, we calculated the mean fractional methylation value for all the mapped

cytosines within the CpG island. For each mapped cytosine, the fractional methylation value was calculated as: total number of "C" reads / (total number of "C" reads + total number of "T" reads), following previous studies [92]. To identify differentially methylated CpG islands between oocyte and sperm in mouse, we extracted the location of these CpG islands from, and used the UCSC liftOver tools to convert the coordinates from genome build mm10 to hg18.

Genomic control regions (non-CpG island controls) are obtained using the following procedures. We first removed all the CpG islands from the whole genome and then randomly sampled genomic regions with the length distribution identical to the list of CpG islands. Methylation levels of these control regions are calculated the same way as CpG islands. We repeated this procedure 1000 times. We performed this sampling with and without removing CpG islands. The results from both analyses are consistent, and in this dissertation we present results obtained without removing CpG islands.

Identification of High Methylation Variability Regions

We identified highly differentially methylated regions ('High Methylation Variability Regions, HMVRs') from the comparison of the five methylomes using a sliding window approach. First, we calculated coefficient of variation of DNA methylation (standard deviation/mean) of all CpG sites in the human genome, using the comprehensive methylation information of all nucleotides in the five methylomes. Then, we identified individual CpG sites that exhibit high variation using a cutoff value of C.V. = 1 (this roughly corresponds to >2.4SD, and top 0.5% quantile of the whole CpG sites). A total of 0.69 million CpGs out of 21.7 millions examined CpGs sufficed this criterion. Then, using a 2-kb sliding window with a 200 bps step size of increment, we extended the window until each window contained less than 50% of HMV CpG sites. HMVRs are then defined as genomic regions that include 5 or more HMV CpG sites. We also tried

combinations of different parameters to identify HMVRs: our results did not change qualitatively.

Hierarchical Clustering Analyses

Clustering of CpG islands of the five tissues methylome data was performed using a function called 'clustergram' in MATLAB. It employs hierarchical clustering with a Euclidean distance metric to first cluster the tissues and then cluster the CpG islands. Ward linkage was employed to generate both dendograms.

Analyses of Gene Expression

Gene expression data from 6 human tissues (prefrontal cortex, cerebellum, heart, kidney, liver and testis) were based upon whole genome RNA sequencing [93]. These data were aligned to the respective genome sequences by the TopHat program. The expression levels were normalized by mean per-base read coverage with unambiguously mapping reads. The samples measured for the same tissue were averaged to represent the expression level for that specific tissue. The second data set is based upon Affymetrix human genome U133A array which were downloaded from Gene AtlasV2 (GSE1133), where the expression level is standardized by MAS5.0 algorithm [94]. We removed disease tissues and used only normal tissues. Based upon these expression values, the "tissue specificity index" [95] is defined by incorporating information on the maximum expression level among the tissues in each data set, as follows:

$$T = \frac{\sum_{j=1}^{n} \left(1 - \left[\log_2(E_j) / \log_2(E_{\max})\right]\right)}{n-1}$$

where n is the number of tissues analyzed, E_j the expression level of the gene in the jth tissue and E_{max} the maximum expression level of the gene across the 6 tissues. The higher the tissue specificity index of a gene, the more the tissue-specific its expression pattern is. To define the association of the genes to certain CpG island, we first extended

the genes at both 5' and 3' end by 1500 bp, and we called that the gene is associated with certain CpG island if there is any overlap between the extended region and the CpG island.

To examine overlaps between CpG islands and transcription factor binding sites, we downloaded the location of transcription factor binding sites conserved in the human/mouse/rat alignment from UCSC genome browser. A binding site is considered conserved across the alignment based upon the score threshold computed with the Transfac Matrix Database (v7.0). Transcription factor binding sites that are completely located within the CpG islands are counted for each CpG island.

Evolutionary Substitution Rates of CpG Islands

We used Cohen et al. [85]'s evolutionary data downloaded from the Tanay lab website (http://compgenomics.weizmann.ac.il/tanay). The data consist of a list of bigWig tracks containing observed and expected evolutionary dynamics in 50bp resolution, smoothed over 2kb windows. We converted the bigWig encrypted files to bedGraph files using the UCSC utility bigWigToBedGraph. We then computed the weighted average of observed and expected rates for each CpG island region using custom perl scripts.

Discriminant and Classification Analyses

We performed linear discriminant analyses using the "lda" function from the package of "MASS" in R. We also performed support vector machine analyses using the "ksvm" function from the package of "kernlab" in R. For both analyses, 20% of the whole data were randomly selected as the training data set. After training the model, the predictions were made for the test data set and the accuracy was evaluated based upon the comparison between prediction and the actual label in test data set.

RESULTS

Patterns of CpG Islands DNA Methylation in Whole Genome Methylomes

We analyzed comprehensive whole genome nucleotide-resolution DNA methylation maps (referred to as 'methylomes' henceforth) from eight distinctive human samples (prefrontal cortex of brain, embryonic stem cells, neonatal foreskin fibroblasts, peripheral-blood mononuclear cells, sperms, placenta, cerebellum, and kidney: Table 2.1). We did not include methylomes originating from induced pluripotent stem cells [96], as the epigenetic patterns of these cells may differ from those from normal somatic cells. Among these methylomes, five (prefrontal cortex of brain, embryonic stem cells, neonatal foreskin fibroblasts, peripheral-blood mononuclear cells, sperms) methylomes offer similarly comprehensive, high coverage information across the whole genome (Table 2.1). Including data from all eight methylomes reduced the number of examined CpG sites dramatically (~8 fold: Table 2.1). Nevertheless, the results of the analyses of these five methylomes and the total eight methylomes are highly similar, and for the rest of the results in this study, only results from the five comprehensive methylomes are presented.

We calculated methylome-specific DNA methylation levels of CpGs across the whole genome (Materials and Methods). From the five whole genome methylomes, we annotated methylome-specific DNA methylation of 26.7 million CpG dinucleotides, corresponding to 88.7% of all CpG dinucleotides in the human genome. Using this method, we determined methylome-specific levels of DNA methylation for 25,131 CpG islands, encompassing 89% of all annotated CpG islands in the UCSC genome browser. Comparisons to methylation data from other methods indicate that the data we used offer superior resolution for examining the detailed variation of DNA methylation in CpG islands (Figure 2.1).

As expected, CpG islands exhibit significantly reduced methylation compared to the genomic background (Mann-Whitney test, $P < 10^{-15}$, Figure 2.2A). Notably, CpG

islands in sperms are the least methylated among the 5 methylomes, even though the sperms themselves are not the least methylated among those (Figure 2.2A). This is not due to inactivation of the X chromosome during spermatogenesis: this pattern persists even when the data from the X chromosome is removed. When examined individually, the majority of CpG islands are hypo-methylated (methylation level <20%). However, substantial numbers of CpG islands are hyper-methylated (methylation level >80%) (Figure 2.2B). The percentages of hyper-methylated CpG islands range between 15% in sperm, to 23% in embryonic stem cells. Interestingly, we discovered a strong negative correlation between CpG island lengths and the average methylation levels across the methylomes: longer CpG islands tend to be more markedly hypo-methylated (Spearman's $\rho = -0.38$, $P < 10^{-16}$, Figure 2.2C).

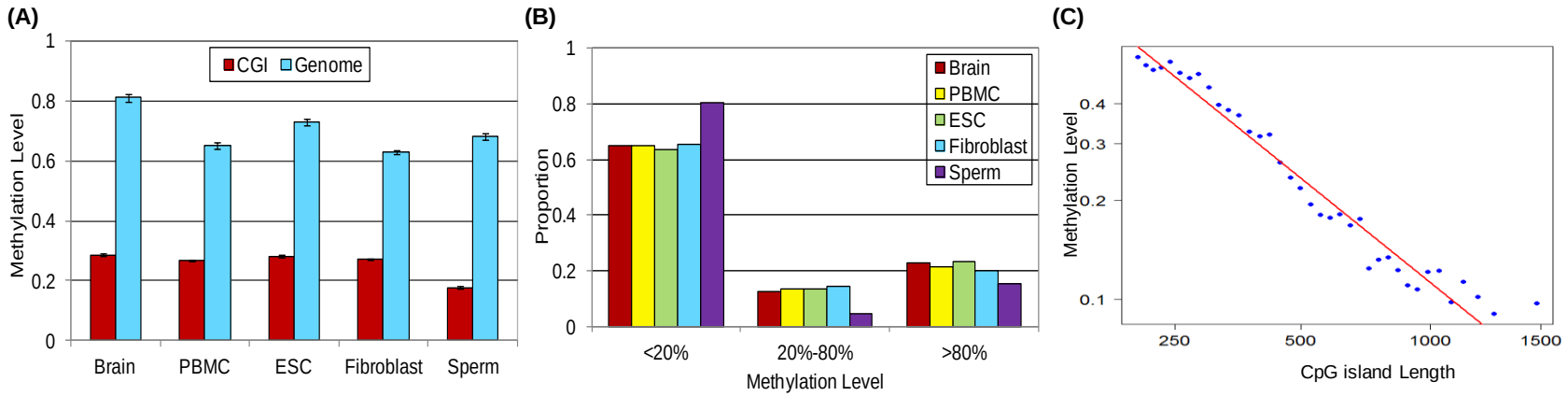

Figure 2.2. Overview of DNA methylation at CpG islands across tissues (A) Mean methylation levels of CpG sites in genomic background (blue bars) versus those in CpG island context (red bars). (B) Distribution of CpG islands that are lowly methylated (<20% mean fractional methylation levels), intermediately methylated (20%~80% mean fractional methylation levels) and highly methylated (>80% mean fractional methylation levels) across the five methylomes examined. (C) Correlation of CpG island length and methylation level. A regression of log transformed CpG island length versus log transformed average methylation level from 5 human methylomes, divided into 40 bins, shows a high negative correlation (R^2 = 0.96 for binned data).

CpG Islands Mark Highly Variable DNA Methylation Regions

Having demonstrated the overall hypo-methylation of CpG islands and hyper-methylation of some CpG islands, we then examined DNA methylation variability of CpG islands. Several studies have identified differentially methylated CpG dinucleotides and genomic regions among different tissues and cell types, many of them (but not all) included CpG islands [97-103]. Analyses of whole genome methylation maps provide a unique opportunity to identify genomic regions whose methylation levels vary between different tissues. With the methylation level of individual CpGs present in 5 methylomes, we examined variability of DNA methylation. For this purpose, we used the coefficient of variation (CV: standard deviation divided by mean), which is a commonly used metric to compare the level of variability of biological data [104-106]. Then we developed a sliding window approach to define 'high methylation variability regions' (HMVR, Materials and Method) of the human genome. We identified a total of 17,045 HMVRs, spanning 51.2 million bps and containing 0.70 million CpG dinucleotides. Remarkably, CpG islands are highly significantly over-represented in these HMVRs. Under a criterion of over 80% overlap, 12,683 CpG islands overlap with these HMVRs. In comparison, the expected number of CpG islands in HMVRs is only 483 ($P < 10^{-20}$, Fisher's exact test). Similar results were obtained when different criteria were used to identify HMVRs. Thus, even though generally hypo-methylated, CpG islands in fact exhibit tremendous level of DNA methylation variation across methylomes. This pattern is apparent when we compare the variability of methylation levels of CpG islands to those of 'control' regions (Figure 2.3).

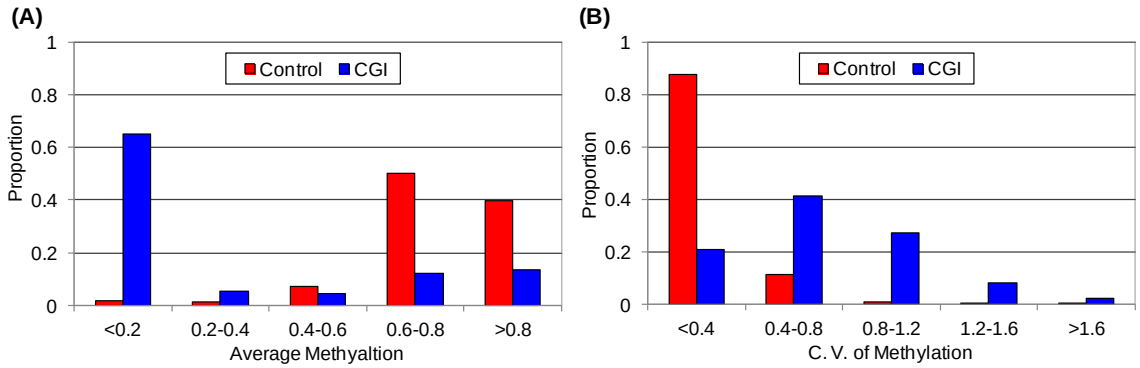

Figure 2.3. Comparisons DNA methylation level and variation between CpG islands and control region (A) CpG islands (blue bars) exhibit lower level of DNA methylation compared to genomic background (red bars). (B) Levels of DNA methylation variation, measured by C. V. (coefficient of variation), are higher for CpG islands than genomic background.

Distinctive Clusters of Human CpG Islands Based On DNA Methylation Patterns

Having established that CpG islands exhibit highly variable DNA methylation across the whole genome methylomes, we investigated patterns of DNA methylation variation more deeply. We employed a hierarchical clustering approach (Materials and Methods) to group CpG islands according to their similarities of DNA methylation across the five methylomes. The resulting 'heat map' of DNA methylation variation across CpG islands reveals several intriguing patterns (Figure 2.4). Interestingly, the clustering pattern of the methylomes does not reflect the gender or developmental stages of the original tissue or cell samples: among the five methylomes, only the ESC has female origin. ESC methylome clusters with other methylomes of male origin. It is also notable that the ESC methylome is closer to those of highly differentiated cell methylomes. On the other hand, sperm methylome is the most distinct among the five methylomes, which is consistent with our previous results indicating that sperm CpG islands are the least methylated among the five methylomes. This pattern highlights epigenetic differences between germ lines and somatic tissues, and regulatory effects of DNA methylation on spermatogenesis [107-109]. CpG islands in embryonic stem cells also exhibit distinct

patterns of DNA methylation from that of other tissues, highlighting the unique developmental potential of these cells.

Strikingly, CpG islands form several distinct clusters according to their methylome-specific DNA methylation patterns (Figure 2.4). As expected, many CpG islands exhibit sparse levels of DNA methylation in all five methylomes. These are designated as 'Cluster I' (Figure 2.4A). Note that many CpG islands in this cluster still exhibit high levels of methylation variability (Figure 2.4A). The rest of CpG islands are differentially methylated in different methylomes. Among these, approximately half of CpG islands are notably hypo-methylated in sperms, yet exhibit highly variable patterns of methylation in somatic tissues and embryonic stem cells (cluster II, Figure 2.4A). The remaining CpG islands tend to be heavily methylated in sperms and methylated in some somatic tissues and embryonic stem cells (cluster III in Figure 2.4A). We can further divide the clusters II and III to sub-clusters, which exhibit distinctive variability of DNA methylation. For example, cluster II can be subdivided into those that exhibit sparse methylation in sperm but relatively heavy methylation in somatic cells (sub-cluster IIa), and those exhibiting sparse methylation in sperm and highly variable and often sparse patterns of methylation in somatic cells (sub-cluster IIb) (Figure 2.4B). Cluster III includes a distinctive subcluster of CpG islands that exhibit heavy methylation in all tissues (sub-cluster IIIa), compared to those that show variably methylated across tissues (sub-cluster IIIb, Figure 2.4C).

To determine whether the observed pattern is applicable to a larger number of tissues and cell types, we incorporated nucleotide-resolution DNA methylation maps from additional three methylomes generated from placenta, cerebellum and kidney (Materials and Methods). These additional methylomes consist of markedly lower sequencing coverage and/or few CpG sites compared to the five comprehensive methylomes. Despite such difference in sequence coverage and quantity, clustering analyses using these eight methylomes clearly demonstrate distinctive CpG islands

clusters that are highly similar to the above results (Figure 2.5). Together, these results indicate that human CpG islands can be clearly classified into several groups according to the patterns of DNA methylation variability across multiple methylomes.

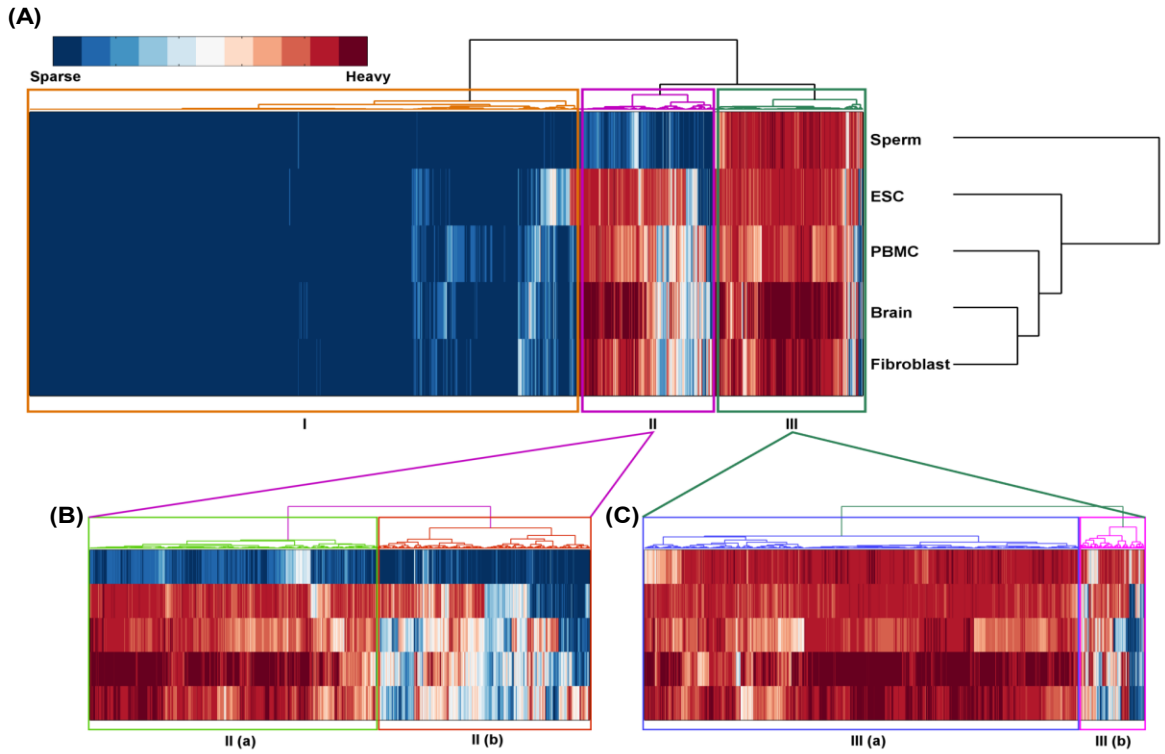

Figure 2.4. Hierarchical Clustering of CpG islands according to their methylation levels in 5 human methylomes. The bar on top left represents relative methylation levels, where "Heavy" stands for the methylation level of 100% while "Sparse" stands for no methylation. (A) Three distinctive clusters are indicated (B) Some CpG islands are hypo-methylated in the sperm methylome, but hyper-methylated in other methylomes (IIa, n=2357) or exhibit variable levels of hyper-methylation in other methylomes (IIb, n=1751). (C) Some CpG islands are generally hyper-methylated in all methylomes (IIIa, n=3885) or exhibit some level of tissue-specific hypo-methylation (IIIb, n=589).

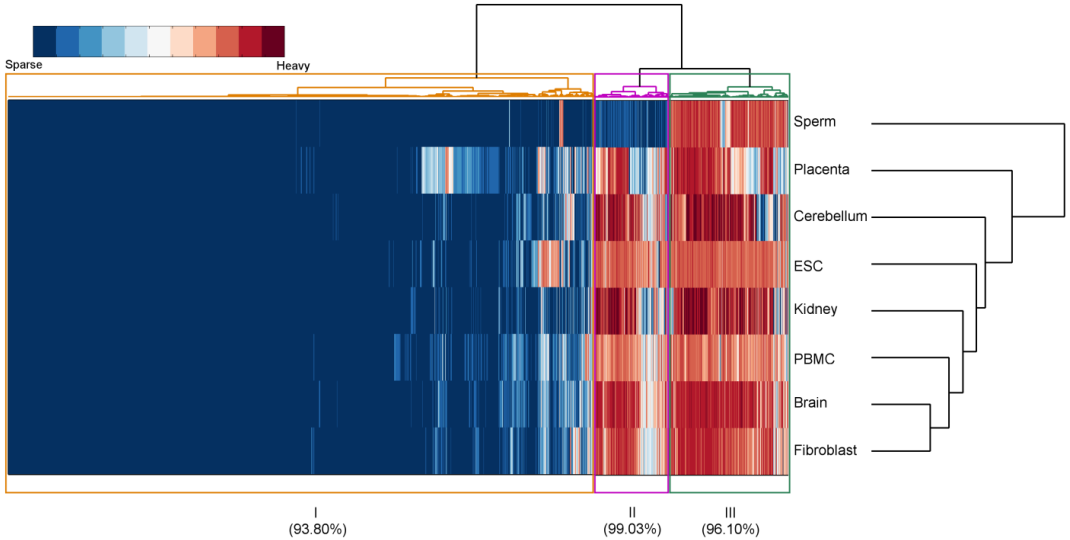

Figure 2.5. Hierarchical Clustering of CpG islands according to their methylation levels in 8 human methylomes. The bar on the top left represents relative methylation levels. Three distinctive clusters are indicated. Cluster I , II, II consists of 11231, 1437 and 2283 CpG islands respectively. Individual cluster consistency percentage with 5-tissues clustering is indicated in the parenthesis below each cluster numbers. The total consistency is 94.66% .

Epigenomically Identified CpG Island Clusters Are Genomically Distinctive

Having identified distinctive patterns of DNA methylation variability across CpG islands, we examined whether these CpG islands clusters exhibit different characteristics. Intriguingly, we find that these clusters, which have been identified solely based upon patterns of DNA methylation variation, differ significantly in several genomic characteristics. The cluster I CpG islands tend to be the longest, which is consistent with our observation that longer CpG islands tend to be less methylated. They also contain the most G and C nucleotides and exhibit the highest CpG O/Es, and harbor the largest numbers of CpG dinucleotides, compared to those in other clusters (Figure 2.6A-C). On the other hand, cluster III CpG islands are distinctively shorter than those in other clusters, as well as exhibiting lower GC contents and lower CpG O/Es. Notably, these CpG islands consist of strikingly fewer numbers of CpG dinucleotides compared to those in other clusters (Figure 2.6D). In comparison, CpG islands in the cluster II generally exhibit genomic characteristics that are intermediate of the other two clusters. These differences are not due to the bias in mapping: CpG islands in the three clusters show similarly high mapping coverages. Autosomal and X-linked CpG islands also exhibit heterogeneous distribution: CpG islands on the X chromosome are slightly yet significantly enriched in cluster I, while deficient in cluster III (Table 2.2). These clusters also exhibit enrichment of distinctive genomic regions: cluster I consist of largely promoter-associated CpG islands, while clusters II and III include large numbers of intragenic and intergenic CpG islands (Figure 2.6E). The observation that CpG islands in clusters II and III tend to exhibit highly methylome-specific patterns of DNA methylation is thus consistent with the idea that intragenic and intergenic CpG sites are highly variably methylated and exhibit tissue- specific DNA methylation [110].

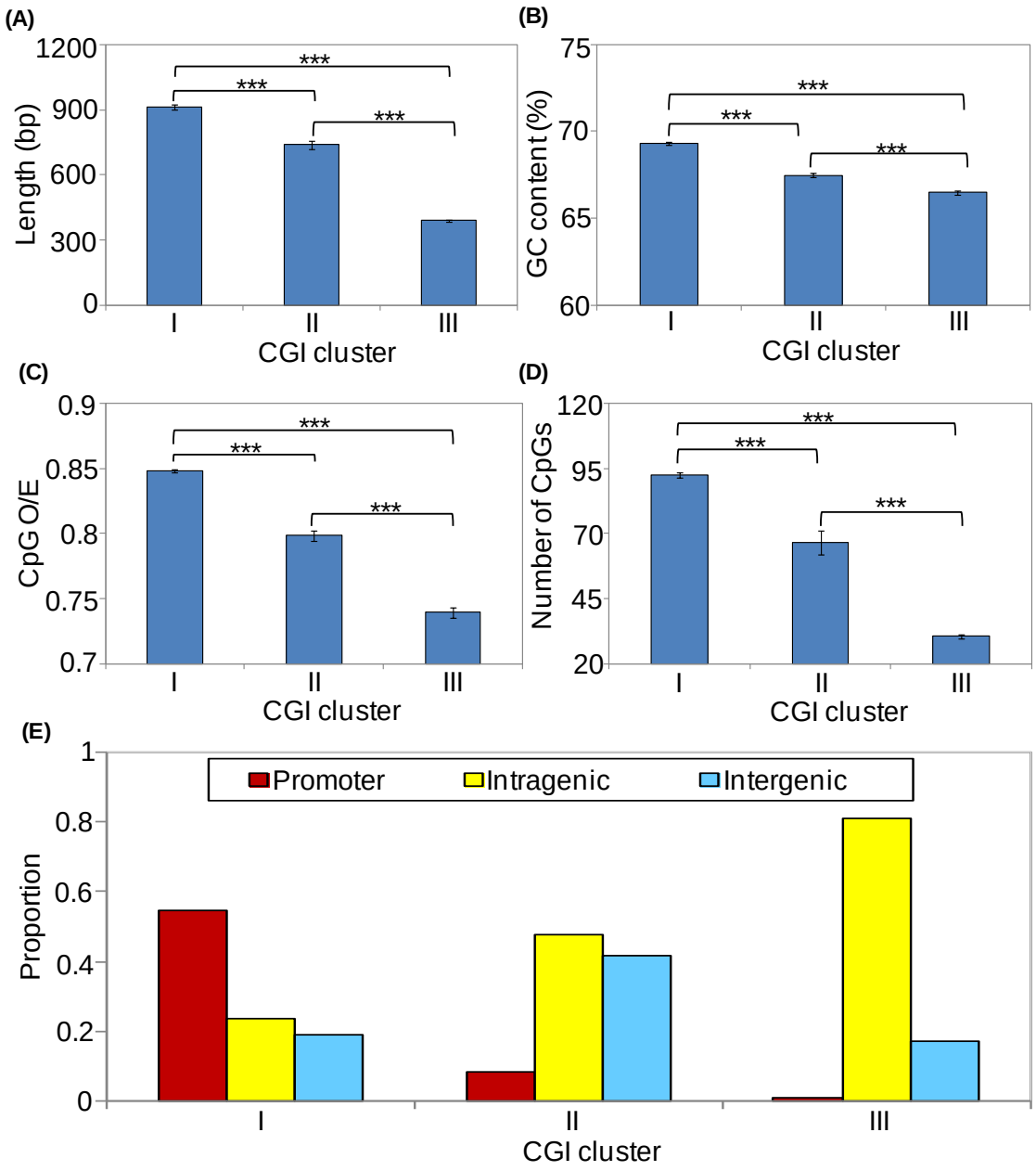

Figure 2.6. Contrasting genomic features of the three CpG island clusters. Significant differences are found in (A) lengths, (B) GC content, (C) CpG O/E, and (D) number of CpG dinucleotides among the three clusters. (E) Occurrence of promoter-, intragenic- and intergenic- CpG islands across the three CpG islands clusters. Significance level ***: $P < 10^{-6}$.

Table 2.2. The numbers of X-linked CpG islands compared to all CpG islands. X-linked CpG islands are over-represented in the cluster I and under-represented in cluster III.

	Cluster I	Cluster II	Cluster III	
X-linked CpG islands	366	76	27	
All CpG islands	16183	4032	4447	
				$P < 10^{-16*}$

DNA Methylation Variation Supports Evolutionary Diversity of CpG Islands

As genomic features are determined by evolutionary processes, we sought to determine underlying evolutionary mechanisms of distinctive CpG island clusters. For this we used recently available evolutionary classification of CpG islands by Cohen et al. [85]. This study used parameter-rich evolutionary models to infer evolutionary forces underlying the evolution and maintenance of CpG islands in primate genomes. Because methylated cytosines frequently mutate to thymines, DNA methylation in effect depletes CpG dinucleotides [111, 112]. It was originally proposed that CpG islands manage to maintain high CpG contents against this mutational pressure of DNA methylation by avoiding DNA methylation [67-69]. Indeed, Cohen et al. identified many CpG islands with evolutionary signatures of hypo-methylation. They named these CpG islands as 'hypo-deamination' CpG islands. On the other hand, their analyses revealed that some CpG islands maintain CpG contents via biased gene conversion process (referred to as 'biased gene conversion' CpG islands). They also identified 'pseudo' CpG islands, which are genomic regions that happen to harbor large numbers of CpG dinucleotides by chance, and expected to lose their CpG contents through evolution.

Our data on high resolution DNA methylation variation of CpG islands provide a novel way to test some of the predictions and implications of these evolutionary analyses. First, 'hypo-deaminated' CpG islands should exhibit hypo-methylation in germlines. We found that cluster I CpG islands, which are hypo-methylated in all five methylomes, are

overrepresented in 'hypo-deamination' groups (Figure 2.7B). The fact that they are also hypo-methylated across somatic tissues (in addition to in sperms) indicates that evolutionary pressures for hypo-methylation of these CpG islands may share some of the same underlying mechanisms with somatic hypo-methylation. On the other hand, we detected a strong influence of biased gene conversion in CpG islands in clusters II and III (Figures 2.7C and 2.7D). Because biased gene conversion process preferentially fixes C and G nucleotides, they can generate CpG dinucleotides, and consequently, counter-balance the mutational depletion of CpGs. CpG islands in the cluster III also include disproportionately large numbers of 'pseudo' CpG islands (9%, significantly higher than 0 and 2% in CpG island clusters I and II).

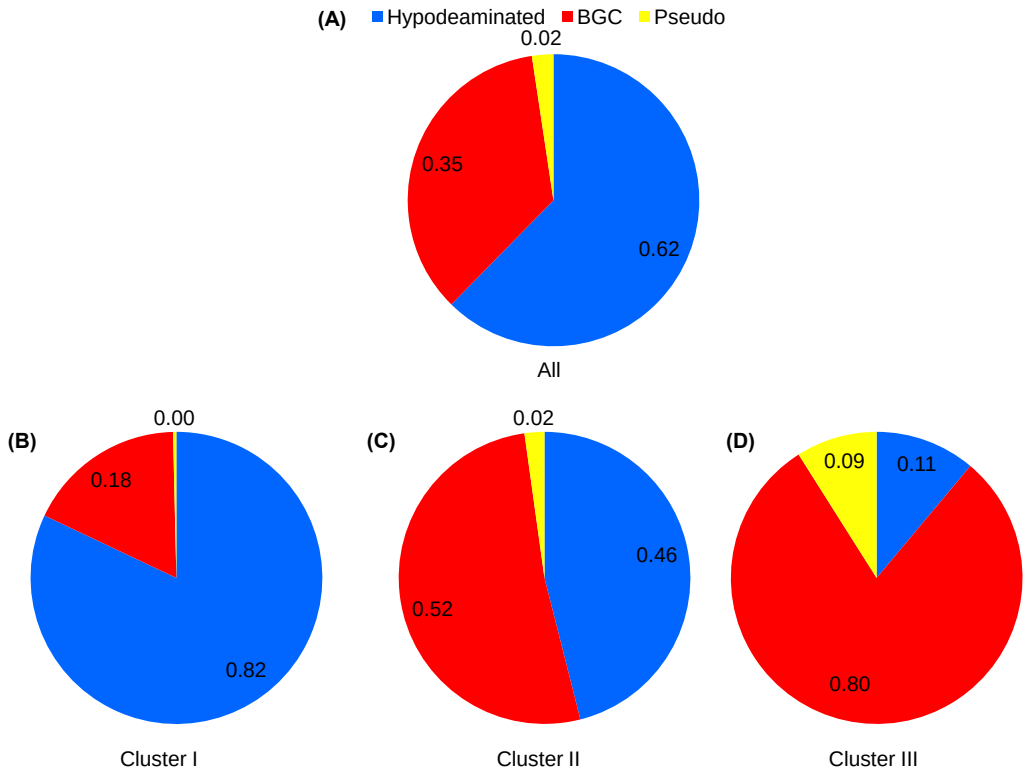

Figure 2.7. Evolutionary Diversity of 3 CpG Island Clusters. Frequencies of hypo-deaminated, biased gene conversion (BGC), and pseudo CpG islands in (A) all CpG islands, (B) cluster I, (C) cluster II and (D) cluster III.

Functional Diversity of CpG Islands Reflected in DNA Methylation Variation

We have so far demonstrated that epigenetic variability of CpG islands represents distinctive clusters, and is tightly linked to genomic and evolutionary variability in the human genome. We posit that these distinctive CpG clusters further indicate regulatory diversity among CpG islands. Specifically we hypothesize that the regulatory functions of cluster I CpG islands are tightly linked to their hypo-methylation status. In contrast the regulatory effects of CpG islands in the clusters II and III may critically rest on their cell- and tissue-specific hypo- and hyper-methylation.

To test these hypotheses we performed several analyses to examine functional diversity among CpG island clusters in the human genome. First, we performed gene ontology analysis to determine whether these clusters are enriched in different functions. Indeed, these clusters are enriched in highly functionally distinct genes (Table 2.3). Cluster I CpG islands are generally associated with genes participating in 'housekeeping' functions such as transcription and RNA-processing. In addition, some developmental functions, in particular neuron development, are also overrepresented in Cluster I. Cluster II CpG islands are associated with genes involved in morphogenesis and cell-cell adhesion. Genes associated with cluster III CpG islands have fewer ontology terms that are significantly enriched, which include protein phosphorylation, negative-regulation pathways, and signal transduction (Table 2.3).

These results are consistent with the idea that hypo-methylation of Cluster I CpG islands may regulate housekeeping functions, while variable DNA methylation of Clusters II and III may regulate tissue- and developmental stage- specific functions. We further hypothesize that these distinctive functions are likely to be achieved by distinctive transcriptional regulation. Specifically we hypothesize that the cluster I may represent generally highly expressed genes across tissues and cells while the clusters II and III encode more variably expressed genes. To test this hypothesis, we examined tissue-specific transcription of associated genes utilizing recent RNA-seq based gene expression

profiles from six distinct human tissues [93]. Tissue specific pattern of expression is summarized by the "tissue specificity index" measure (Materials and Methods). We found that genes associated with the cluster I CpG islands are the most broadly expressed (tissue specificity is the lowest) compared to those associated with cluster II and III CpG islands (Figure 2.8B). Genes associated with the cluster II exhibit the most tissue-specific patterns of gene expression (Figure 2.8B). Cluster III CpG islands are associated with genes demonstrating intermediate tissue specificity of gene expression compared to the other two clusters. To ascertain that this observation is consistent across large number of different types of tissues, we also analyzed the Novartis tissue specific gene expression data, encompassing 67 normal tissues [94]. Analyses of these data again indicate that the genes associated with the CpG island cluster I are most broadly expressed and the genes associated with clusters II and III are associated with narrower gene expression patterns (Figures 2.8A,B). Thus, tissue- specific DNA methylation of CpG islands may contribute to tissue- specific expression of associated genes.

One way in which CpG islands affect transcriptional regulation is by encoding transcription factor binding sites [113]. We examined how often CpG islands overlap with transcription factor binding sites conserved in the human/mouse/rat alignment (Materials and Methods). We found that the average number of transcription factor binding sites, after normalized by lengths, is significantly higher in CpG islands than in the control regions. Interestingly, the cluster I CpG islands has the largest number of TFBSs while the cluster II has the least (Figure 2.8C). This is consistent with the observation that the cluster I is enriched in promoters, while the others are often found in intergenic and intragenic regions (Figure 2.6E). It is also consistent with experimental results demonstrating that ubiquitously active promoters harbor large numbers of transcription factor binding sites and many CpGs, while promoters that are tissue-specific have fewer CpGs [114]. At the same time, even tissue-specific CpG islands appear to encode large number of potential transcription factor binding sites (Figure 2.8C).

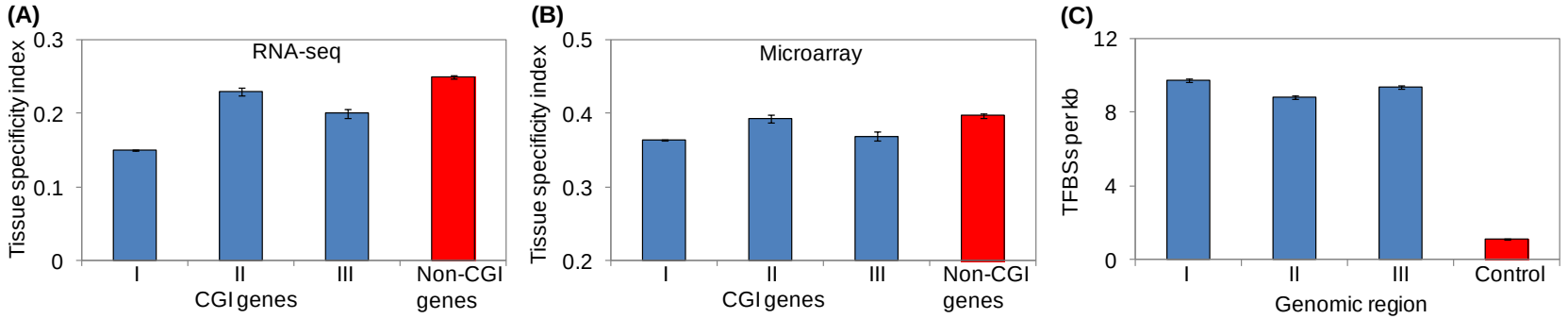

Figure 2.8. Contrasting expression patterns and transcription factor binding sites of the three CpG island clusters. Tissue specificity gene expression indices based upon RNA-seq (A) and microarray (B) data are shown for the CpG islands genes (blue bars) and non CpG islands genes (red bar). (C) The mean numbers of TFBSs (per kb) for each CpG island cluster (blue bars) and control regions (red bar)

Table 2.3. Distinctive functional enrichments of specific genes according to the variable DNA methylation of CpG islands.

GO terms	Description	*P*-values	*FDR-P*-values*
Cluster I CpG islands			
sparse sperm methylation, sparse ESC and somatic cell methylation			
GO:0006350	transcription	2.00×10^{-28}	3.90×10^{-25}
GO:0045449	regulation of transcription	3.77×10^{-27}	7.36×10^{-24}
GO:0006396	RNA processing	1.11×10^{-17}	2.16×10^{-14}
GO:0030182	neuron differentiation	2.49×10^{-15}	4.76×10^{-12}
GO:0051252	regulation of RNA metabolic process	5.14×10^{-15}	9.96×10^{-12}
Cluster II CpG islands			
sparse sperm methylation, variable ESC and somatic cell methylation			
GO:0007156	homophilic cell adhesion	1.64×10^{-9}	3.00×10^{-6}
GO:0016339	calcium-dependent cell-cell adhesion	9.42×10^{-9}	1.72×10^{-5}
GO:0007155	cell adhesion	1.83×10^{-8}	3.35×10^{-5}
GO:0022610	biological adhesion	1.98×10^{-8}	3.62×10^{-5}
GO:0048598	embryonic morphogenesis	6.95×10^{-7}	1.27×10^{-3}
Cluster III CpG islands			
variable methylation in all five methylomes			
GO:0006468	protein amino acid phosphorylation	6.89×10^{-6}	0.013
GO:0051056	regulation of small GTPase mediated signal transduction	1.06×10^{-5}	0.019
GO:0031327	negative regulation of cellular biosynthetic process	1.40×10^{-5}	0.025
GO:0007010	cytoskeleton organization	2.62×10^{-5}	0.048
GO:0046578	regulation of Ras protein signal transduction	2.74×10^{-5}	0.050

CpG Islands in Disease, Genomic Imprinting, and Aging

Having established the CpG island clusters and their distinctive properties at genomic, evolutionary and functional levels, we further hypothesized that CpG islands with respect to different aspects of human healths and aging may show different enrichment over the three clusters. We first asked whether CpG islands in certain clusters tend to be over-represented in disease, in particular cancer. A recent study [115] compared DNA methylation maps of over 1,149 tumors of different tissue origins and identified genes whose CpG island promoters frequently exhibit aberrant hyper-methylation in cancers. Among these promoters that are prone to aberrant hyper-methylation in cancers, 663 overlapped with our CpG island data. We find that 649 (97.8%) of them belonged to cluster I, an extremely significant over-representation ($P < 10^{-20}$, Fisher's exact test). The remaining 14 CpG islands are from the cluster II.

We next examined the association between imprinted genes and different CpG island clusters. Imprinted regions are expected to be differentially methylated between germline and somatic cells. To test these hypotheses, we downloaded a list of monoallelicly expressed human genes from the genomic imprinting website (http://www.geneimprint.org/, Materials and Methods). Among these imprinted genes, 33 overlapped with the CpG islands in our data. Thirteen out of these 33 imprinted genes are found in the cluster II, representing a significant enrichment (the expected number of imprinted genes in the cluster II is 5, $P < 0.05$ by Fisher's exact test; Figure 2.9). Thus, as expected, imprinted genes are over-represented in cluster II which is distinctively methylated between germlines and somatic cells.

In addition we investigated whether CpG islands that exhibit differential DNA methylation with respect to aging tend to be preferentially associated with specific clusters. Recently, whole genome DNA methylation maps of three individuals of different ages (newborns, 26 years old, and a centenarian) have become available [116]. This study has identified 17,930 'aging' differentially methylated regions (DMRs). The

occurrences of 294 CpG islands overlapping with these aging-DMRs in the three clusters are shown in Figure 2.9. While these CpG islands are distributed across all three clusters, they are highly significantly over-represented in the cluster II, and significantly under-represented in the clusters I and III (Figure 2.9).

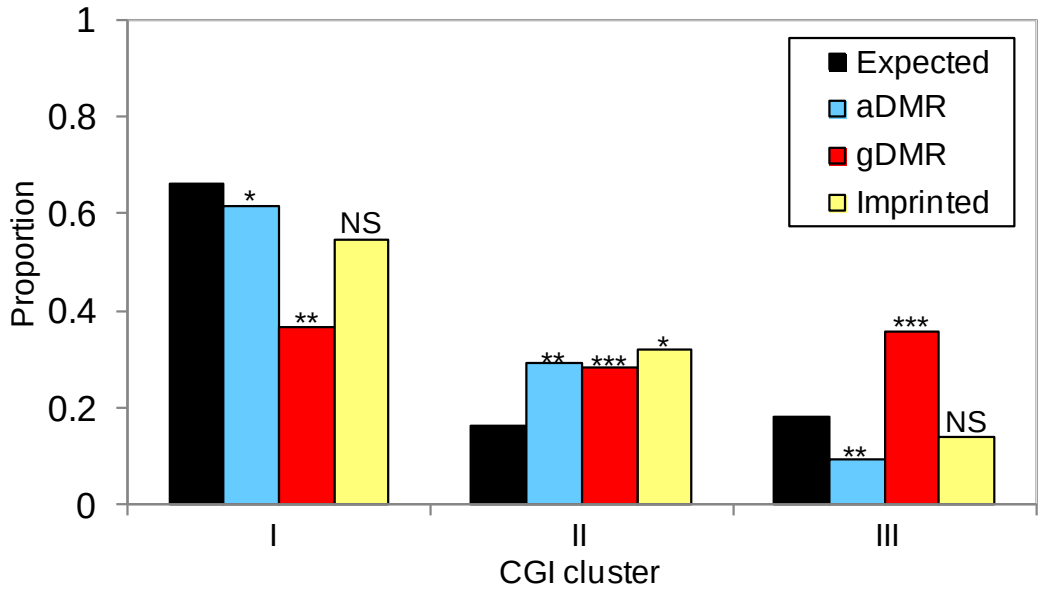

Figure 2.9. Non-random association between CpG island clusters and distinctive biological processes, including gene imprinting (Imprinted), differentially methylation region associated with aging (aDMR) and germ line (gDMR). Significance is assessed by Fisher's exact test: NS $P > 0.05$; *$P < 0.05$; ** $P < 10^{-6}$; *** $P < 10^{-9}$.

DISCUSSION

CpG islands are considered as genomic regulatory hotspots. The advent of molecular techniques to examine nucleotide level DNA methylation of all CpG dinucleotides in a genome provides an exciting opportunity to investigate detailed variation of DNA methylation of these important regulatory regions. Here we examined variation of DNA methylation across multiple methylomes of distinctive origins. We showed that on average CpG islands are highly hypo-methylated, consistent with the prevailing idea that CpG islands generally lack DNA methylation [68, 69]. However, we also showed that many CpG islands exhibit highly variable patterns of DNA methylation

across multiple methylomes (Figure 2.3). In fact, CpG islands exhibit significantly greater degrees of methylation variation than genomic background (Figure 2.3), a surprising pattern counter to the general notion that DNA methylation levels of CpG islands are stable[117]. CpG islands are also more highly variable than adjacent regions, or 'CpG shores'[101] in our data (Figure 2.10). When we examined methylation levels of CpG island shores (defined as 2kb upstream of CpG islands), CpG island shores exhibited lower variability of DNA methylation compared to CpG islands (Figure 2.10). We further examined CpG sites that are ranked at the top 1% of their variability. These sites are highly enriched with those belonging to CpG islands. Sites classified as CpG shores also exhibit significant enrichments in these sites, but not as strikingly as CpG islands. However, there are several critical differences between our study and the previous studies of CpG shores: it is possible that CpG shores may specifically increase methylation variation in cancer cell lines. In addition, our study does not consider inter-individual variability, which may be an important source of epigenetic variability. However, tissue specific epigenetic patterns such as DNA methylation are conserved between distantly related species such as humans and mouse, despite tens of millions of years of divergence [118-120]. Thus, tissue specific DNA methylation patterns have deeper evolutionary origins than variation due to demographic factors. Future studies of epigenetic variability are necessary to test some of these hypotheses.

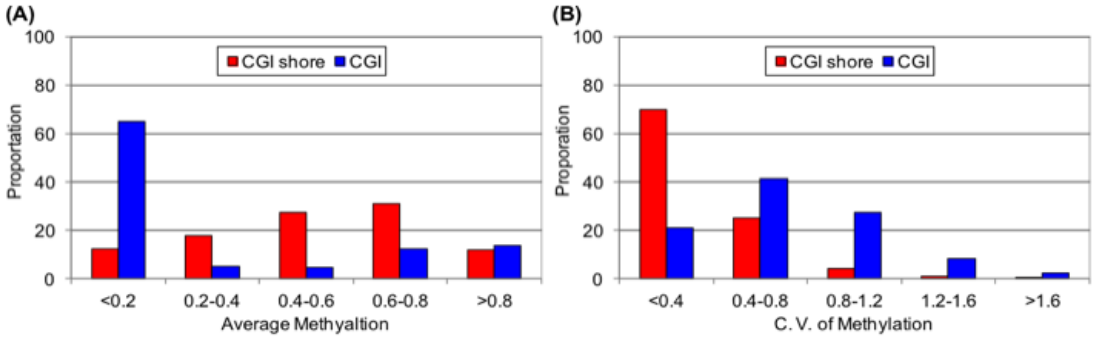

Figure 2.10. Comparison of DNA methylation variability of CpG islands and CpG island shores. (A) Comparison of mean methylation levels of CpG sites in CpG island shores (red bars) versus those in CpG island context (blue bars). (B) Comparison of

methylation variation among 5 human tissues of CpG sites, measured by C. V. (coefficient of variation), in CpG islands versus those in CpG island shores.

Based upon methylome-specific DNA methylation patterns, we can identify several distinctive groups of CpG islands, or 'clusters'. The first cluster of CpG islands are hypo-methylated in all five methylomes. At the genomic level, these CpG islands tend to be longer than those in other clusters, and harbor larger numbers of G and C nucleotides and CpG dinucleotides than others (Figure 2.6). They are also found in or near genes involved in essential, housekeeping pathways including regulation of transcription and RNA processing. These CpG islands generally maintain their CpG contents, and their CpG island status, on an evolutionary timescale by avoiding DNA methylation. Therefore, cluster I CpG islands are closer to the original definition of CpG islands. We refer to them as 'broad' CpG island clusters.

We also found that some CpG islands are sparsely methylated in sperms, yet exhibit variable levels of DNA methylation in other methylomes (Figure 2.4). We tentatively refer to them as 'germline' CpG island clusters. CpG islands in this cluster (cluster II in Figure 2.4A) tend to be shorter and harbor fewer G and C nucleotides and CpG dinucleotides than those in the cluster I (Figure 2.6). They are enriched in cell adhesion and embryonic morphogenesis functions, and generally exhibit more tissue-specific transcription profiles compared to those in the cluster I.

Finally, we show that approximately one fifth of all CpG islands exhibit some degree of DNA methylation in the five methylomes (cluster III in Figure 2.4). They tend to be much shorter than CpG islands in the other two clusters and harbor distinctively fewer CpG dinucleotides (Figure 2.6). The fact that these are generally hyper-methylated in the examined methylomes raises the possibility that these regions may not encode regulatory potentials of true CpG islands. However, an appreciable number (a count of 327) of CpG islands in this cluster exhibit evolutionary signatures of hypo-deamination, indicating that these are maintained as CpG islands by avoiding DNA methylation

(Figure 2.7). It is also notable that there are some CpG islands that clearly exhibit tissue-specific hypo-methylation among this cluster (Figure 3C, sub-cluster IIIb). These observations indicate that at least some of the CpG islands in this cluster likely to harbor true regulatory potential. On the other hand, almost 10% of these CpG islands were even classified as 'pseudo' CpG islands, which are genomic regions possessing large number of CpG dinucleotides by chance (Figure 2.7). We have used linear discriminant analyses and support vector machine (Material and Method) to examine whether we can separate these cluster III 'pseudo' CpG islands from the rest of CpG islands ('normal' CpG islands). However, these analyses did not indicate the presence of specific genomic features of these 'pseudo' CpG islands that are heavily methylated in all methylomes. Future analyses of DNA methylation variation in larger number of tissues and from different developmental stages are necessary to shed light on the functional significance of CpG islands that appear hyper-methylated in the currently examined methylomes.

We sought to explicitly connect the observed variation of CpG islands at epigenomic and functional levels to the underlying evolutionary features. A recent evolutionary analyses classified CpG islands to several groups, namely 'hypo-deamination', 'biased-gene conversion' and 'pseudo' CpG islands [85]. We examined the correspondence between evolutionary classifications and across tissue variability of CpG islands. First, we confirm that many hypo-methylated CpG islands in the cluster I correspond to hypo-deamination CpG islands. Thus for these CpG islands, the evolutionary classification and epigenetic classification correspond fairly well. Interestingly however, many CpG islands in the 'germline' CpG island cluster, which are hypo-methylated in sperm, were classified as 'biased-gene conversion' CpG islands (Figure 2.7C). This is puzzling; if they do indeed avoid DNA methylation in germlines, they should be more enriched in hypo-deamination CpG islands, similar to the CpG islands in cluster I. To resolve this discrepancy, we hypothesize the following two mutually non-exclusive possibilities. First, the sperm methylation patterns may be

distinctive from the methylation patterns during other stages of spermatogenesis that are more subject to evolutionary dynamics. Second, the sperm DNA methylation profiles may be highly different from those in oogenesis.

Currently, whole methylome data for human oocytes are not available. However, data on sperm and oocyte methylation from mice are available. Specifically, a recent study identified 1678 differentially methylated CpG islands between oocyte and sperm in mouse [121]. We thus examined how the orthologous CpG islands of these differentially methylated mouse CpG islands are distributed among the CpG island clusters. We found that differentially methylated CpG islands between gametes are significantly more prevalent in clusters II and III than expected (Fisher's exact test, $P < 2 \times 10^{-16}$ for both clusters), while they are significantly under-represented in cluster I (Fisher's exact test, $P < 2 \times 10^{-16}$, Figure 2.9). These results provide strong support to the hypothesis that the evolutionary signatures of hypo-deamination in cluster II and III CpG islands is due to differential DNA methylation between sperms and oocytes.

The fact that DNA methylation patterns of human CpG islands do not completely correspond to the evolutionary classifications (Figure 2.7) provides unique insights into the origin and maintenance of CpG islands. Many CpG islands may have arisen by non-methylation related mechanisms such as biased gene conversion (BGC) or by chance (pseudo islands), but have been co-opted as regulatory hotspots due to the epigenetic functional advantage in the current human genome. This pattern is particularly strong for those CpG islands that are variably methylated across different tissues (Figure 2.7).

We show that the variation of DNA methylation across individual CpG islands is complex. Importantly, our analyses illustrate that human CpG islands vary substantially in several biological levels, from genomic and evolutionary features to epigenomic variation, and functional enrichments. These observations suggest the presence of a fundamental diversity underlying regulatory mechanisms of CpG islands. For example, many intragenic and intergenic CpG islands, such as those in the clusters II and III, may

encode cryptic promoters and enhancers that function in a highly tissue- and cell type-specific manner. Furthermore, with the new knowledge gained on the variation of DNA methylation found among in normal cell types and tissues, we can contrast epigenetic variations of clinical interest to those that naturally exist in future studies. For instance, we demonstrated that the majority of cancer-implicated CpG islands are those that belong to the cluster I, which stay hypo-methylated in all examined methylomes so far (cluster I in Figure 2.4). This observation suggests that regulatory function of cluster I CpG islands is tightly linked to their hypo-methylation, and DNA methylation of these CpG islands is likely to be detrimental and particularly disease-prone. On the other hand, 'aging' CpG islands show a distinctive distribution across the three clusters: they are significantly more enriched in the cluster II, which consist of CpG islands whose methylation levels are highly variable in somatic tissues and embryonic stem cells (Figure 2.9). We hypothesize that some of DNA methylation variation observed during the aging process may share common molecular mechanisms with tissue-specific methylation variation. For example some CpG islands may be more prone to stochastic variation of DNA methylation between cell types and with aging. However, it should be noted that many studies of aging DMRs may be confounded by the effect of different cell types present in samples: aging studies often use blood samples which contain diverse cell types, each of them exhibiting potentially different methylation patterns [122, 123]. These effects, particularly the role of immune system related changes of specific cell types on aging-DMRs [124, 125], need to be clarified in future studies.

Our study illustrates that comprehensive epigenetic profiling of distinctive cells will be highly useful in understanding regulatory processes of CpG islands, and consequently, furthering our knowledge on the role of CpG islands in disease and/or aging. Our findings may also have immediate practical implications. For example, the widely used infinium human methylation chip ('Illumina 450K chip') includes 136K positions that are annotated as CpG islands. Compared to the total number of CpG sites

within CpG islands in the human genome (approximately 1.8Mbps), this corresponds to 7.2% of CpG island CpG sites. If we examine the representation of CpG sites with respect to our CpG island clusters, 7.2, 5.8, and 10% of CpG sites belonging to the clusters I, II, and III are targeted by this array. In other words, this array targets significantly a higher proportion of CpG islands in the cluster III, which include substantial number of 'pseudo' CpG islands. Consequently, some of the epigenomic variation detected by these positions may lack true regulatory meanings. On the other hand, given the enrichment of 'aging' CpG islands in cluster II, a method targeting more positions in the cluster II could be more efficient in identifying aging associated variation of CpG islands. Findings from our study may help in designing better methods to examine variation of DNA methylation across different biological conditions.

With the positive outlook on whole genome methylation profiling, we expect that the number of distinctive human methylomes will increase by an order of magnitude within the next few years. We expect our framework of CpG island diversity at many biological levels to be helpful in interpreting such new data. In turn, these new data will allow us to investigate biological diversity of CpG islands more deeply, and answer some of the new questions posited here.

CHAPTER 3

INTRODUCTION

Recombination, which involves the exchange of genetic information by the pairing of homologous chromosomes during meiosis, is a common biological process in diploid eukaryotic organisms [126]. As such, it is a fundamental evolutionary mechanism that profoundly affects genomic variation. For example, more than 50% of the variation in nucleotide heterozygosity across the genome is due to recombination in flies and humans [127]. Evidence also shows that genomic features such as codon bias, nucleotide substitutions and dynamics of repetitive DNA elements are extensively shaped by recombination [128-130]. Due to its essential role in reproduction as well as its critical importance in genetic analyses, major efforts have been dedicated in constructing genome-scale, high-resolution recombination maps [131-133] and developing new molecular techniques to analyze recombination patterns [134, 135]. This work has substantially furthered our understanding of this important biological phenomenon. One important finding is that the distributions of recombination rates are non-uniform across the genomes in many species; recombination events are concentrated in highly localized areas known as "hotspots", which in the human genome are typically 1-2kb long and surrounded by much longer regions that are essentially devoid of recombination [136-138]. For example, in the human genome, about 80% of the recombination take place in less than 15% of the sequence [139]. Moreover, recombination rates appear to vary between individuals, populations and species, indicating that recombination rates can rapidly between closely related species, and even within a species [140-142]. For example, it is proposed, based upon studies of human recombination hotspots, that hotspots can emerge and disappear in as little as 120,000 years, and certainly within the six million years since humans diverged from chimpanzees [143]. Such variable and

heritable patterns also indicate that recombination may evolve in response to natural selection [144, 145].

However, the causative factors underlying this variation are largely unknown. Previous studies have identified numerous factors, including molecular, environmental and demographic factors, that affect recombination rates [146]. At the genomic level, recombination rates are significantly associated with several sequence characteristics such as GC content, gene density, simple repeats, transposable elements and a number of different sequence motifs [147-150]. However, DNA sequence itself does not provide a full explanation for variation of recombination rates. For example, the locations and usage of recombination hotspots vary between extremely closely related species such as humans and chimpanzees [140], and even among human individuals [151], where the underlying sequences are extremely similar. These observations have thus sparked much interest on exploring how epigenetic factors may be involved in determination of recombination patterns.

As an epigenetic modification known to be established at prophase I in meiosis when recombination occurs [152], DNA methylation could be a potential factor affecting meiotic recombination rates. By using methylation-associated single nucleotide polymorphisms (mSNPs) as a surrogate marker for germ line DNA methylation, Sigurdsson et al. [22] reported a significantly positive correlation between recombination rate and DNA methylation. In addition, Auton et al. [153] reported that promoters that have high levels of DNA methylation in human sperms generally exhibit high levels of recombination rates. They also noted that in chimpanzee genomes however the opposite pattern was observed (Auton et al. 2012). Other epigenetic modifications, such as histone modifications, may also affect the location and activity of recombination events. For example, PRDM9, an important *trans*-acting factor that controls hotspots specification in human, contains a PR/SET domain that is capable of trimethylation of histone 3 lysine 4, or H3K4me3 [154]. In mouse genome, H3K4 tri- (H3K4me3) and di-methylation

(H3K4me2), which precedes recombination, are enriched at the *Psmb9* and *Hlx1* hotspots [155], and H3K4me3 in yeast is a prominent and pre-existing marks of active recombination sites [156]. Together, there is substantial amount of support for the hypothesis that epigenetic modifications may affect variation of recombination events. However, what are the main epigenetic modification(s) that may underlie variation of recombination rates, and how do they affect evolutionary dynamics of recombination rates, remain to be resolved. In this study, we explored the impact of epigenetic mechanisms on determining species specific recombination hotspots. We show that at the global level, DNA methylation explains a large amount of variation in recombination rates observed in the human genome. However, DNA methylation levels appear to be a weak indicator of fine scale recombination. On the other hand, specific modifications of histone tails stand out to be important molecular features at the recombination hotspots. We show an extensive overlap between both the H3K4me3 and H3K27me3 enriched regions to the recombination hotspots across the human genome. Together with the elevated recombination rate at the H3K4me3 and H3K27me3 enriched regions, these results indicate that histone modifications may play an important role in shaping the genomic landscape of meiotic recombination in human genome.

MATERIALS AND METHODS

Epigenetic Features

In order to analyze DNA methylation at the genomic level, whole genome, nucleotide-resolution DNA methylation maps (methylomes) generated from prefrontal cortex of human brain [157], and from human and chimpanzee sperms [158] were used. These methylomes were all generated with next-generation bisulfite sequencing technology and have similar number of mapped CpG sites. To estimate methylation levels of specific genomic regions, we calculated the mean fractional methylation value for all the mapped cytosines within that region. For each mapped cytosine, the fractional

methylation value was calculated as: total number of "C" reads / (total number of "C" reads + total number of "T" reads), following the method in the previous chapters. We also downloaded the human sperm profiles of H3K4me3 and H3K27me3, which were generated by the chromatin-immunoprecipitation sequencing method [159]. Enriched genomic regions of these two histone modifications relative to input were identified using the USeq analysis package (http://useq.sourceforge.net), which entails calculating false discovery rates (FDRS) converting from a window binomial p-value.

Sequence Features

For information on GC content, CpG normalized content (CpG O/E), and CpG dinucleotide count, custom Perl scripts were written to search within the human (NCBI 36/ HG 18) and chimpanzee (CGSC2.1/panTro2) genome sequence, downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). We calculated the proportion of repeats from certain genomic region based upon the rmsk table for the location and properties of repeated elements created using the RepeatMasker (http://www.repeatmasker.org), which was built on the Repbase database of repeated elements. For the human genetic map and locations of recombination hotspots, the data were based upon applied statistical inference methods to genome-wide genetic polymorphism data [160], which is the phase II of the International HapMap. The chimpanzee genetic map and recombination hotspots were retrieved from a study using similar methods on polymorphism data of 10 Western chimpanzees [153]. A custom Perl script was used to calculate the recombination rates in certain genomic regions (e.g. recombination hotspots and coldspots). In order to check the overlap between the species-specific recombination hotspots as well as the overlap between the human recombination hotspots and histone modification enriched regions, we intersected the genomic locations by using the liftOver tool from the UCSC Genome Browser. Genomic control regions were obtained using the following procedures. We first removed all the

recombination hotspots from the whole genome and then for each simulation of the control regions, we randomly sampled genomic regions for the number of times equal to the species-specific recombination hotspots and with the identical distribution of CpG number to the species-specific recombination hotspots. We repeated the simulation 1 million times. Methylation levels and recombination rates of these control regions are calculated the same way as recombination hotspots.

Sliding Windows Correlation and Statistical Analysis

The genome-wide analyses were done using three different window sizes (250kb, 500 kb, and 1000 kb). For each window size, we divided the genome into non-overlapping windows. Each window was then assigned values according to its genetic and epigenetic properties (recombination rate, number of CpG dinucleotides, proportion of repeats, GC content, mean fractional methylation level). Prior to multiple linear regression analysis, we first transformed the data to provide a better fit to normal distribution using Box – Cox transformation, a form of lognormal transformation. Linear regression was then done using regional recombination rate as the response variable using a stepwise backward method. All the statistical analyses were done using R package version 2.5.1.

RESULTS

Genome-wide analysis of the correlation between DNA methylation and recombination

In order to test the hypothesis that germline DNA methylation affects rates of meiotic recombination, we analyzed the relationship between experimentally determined sperm DNA methylation levels and recombination rates in 500kb non-overlapping windows across the whole human genome. As a control, we also performed a similar analysis between somatic DNA methylation and recombination using methylation maps

from prefrontal cortex of brain (see Materials and Methods). We observed that recombination rate increases roughly linearly with increasing level of DNA methylation in sperm at the 500kb genomic windows (Pearson's correlation coefficient = 0.211, $P < 10^{-16}$; Figure 3.1A). In contrast, this pattern is not obvious in brain (Figure 3.1B). Recombination rate and DNA methylation level are only slightly correlated in brain (Pearson's correlation coefficient = 0.03, P = 0.01), indicating that the association of DNA methylation and recombination may be unique in germlines. We also performed a genome-wide analysis of correlation between DNA methylation and recombination rate in chimpanzees using the same method. Intriguingly, we found that the trend of the correlation in sperm is weak at most, in the opposite direction to what's observed in the human genome (Pearson's correlation coefficient = -0.04, P = 0.002), and no correlation at all in chimpanzee brain (Pearson's correlation coefficient = -0.002, P = 0.84).

In order to confirm that the observed correlation between DNA methylation and recombination rate in human sperm is robust against different window sizes, we performed the genome-wide analyses using other two different window sizes (250kb and 1000kb). The results show that the significantly positive correlations are present in all analyses with different window sizes (Table 3.1). The correlation coefficients increase with increasing window sizes, implying that DNA methylation may influence recombination rates in a broad-scale.

The observed correlations could be due to other genomic features that influence both DNA methylation and recombination rates [147, 148]. Therefore, we performed a partial correlation analysis accounting for previously suggested sequence factors influencing recombination rate (GC content, repeats) as well as factors correlated with DNA methylation (CpG density). After correcting for these factors, we still found a significant and positive correlation between DNA methylation level and recombination rate (Pearson's correlation coefficient = 0.199, $P < 10^{-16}$). We also built a linear model where recombination rate was a response variable, and sequence features (GC content,

number of CpGs, proportion of repeats) and epigenetic feature (sperm DNA methylation level) were predictor variables in the 500kb genomic windows. We first checked the variance inflation factors (VIFs), which are indicators of multi-colinearity among variables. None of the explanatory variables exhibit VIFs greater than 5 (Table 3.2), demonstrating that we could assess individual contributions of each genomic trait without the influence of multi-colinearity. We then calculated the standardized coefficients, which facilitates an assessment of the strength of association between each predictor variable and the response variable. We found that GC content and proportion of repeats in the genome window are the strongest and second strongest predictors for the recombination rate (Table 3.2). Consistent with the partial correlation results, DNA methylation is also a strong predictor for recombination rates (Table 3.2). In total, the proportion of recombination rate variability explained by the linear model was 0.324.

Table 3.1. Genome-wide analysis of the correlation between DNA methylation and recombination rate using different window sizes.

Window size (kb)	*R*	P-value
250	0.158	$< 2.2 \times 10^{-16}$
500	0.212	$< 2.2 \times 10^{-16}$
1000	0.261	$< 2..2 \times 10^{-16}$

Table 3.2. Multiple linear regression of recombination rate as a response to sequence and epigenetic feature predictors in 500kb genomic window.

	Standardized β	P-value	VIF
GC content	0.375	$< 2.2 \times 10^{-16}$	3.27
Proportion of repeats	-0.237	$< 2.2 \times 10^{-16}$	1.17
Fractional methylation	0.134	$< 2.2 \times 10^{-16}$	1.42
Number of CpGs	-0.044	0.05	3.59
Adjust R^2	0.342		

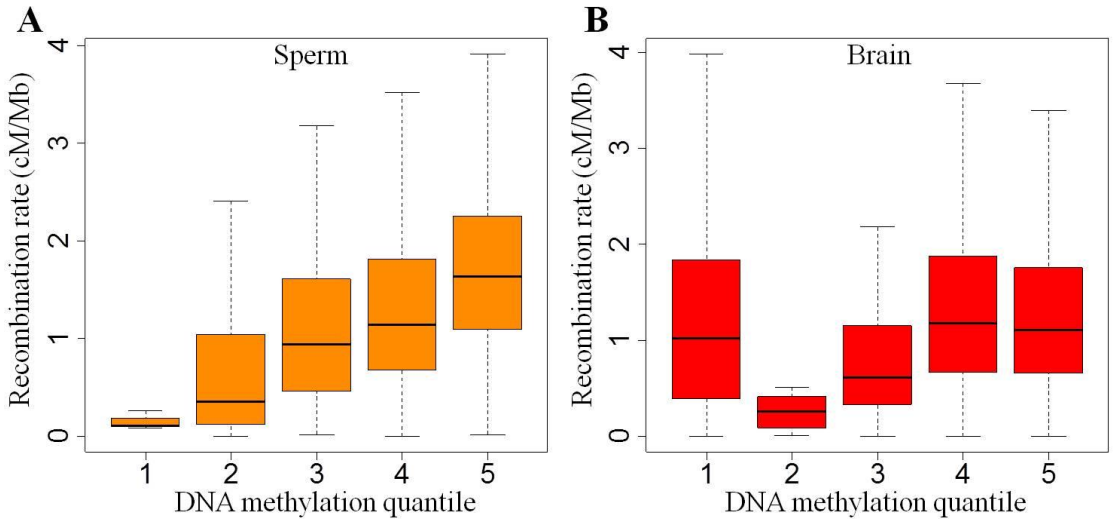

Figure 3.1. Recombination rate is positively correlated with DNA methylation level in sperm but not in brain. Integrating fraction methylation level with recombination rate, we observe a positive correlation between these two factors in sperm (A) but not in Brain (B). The x-axis represents increasing levels of DNA methylation from left to right.

DNA methylation and recombination hotspots

After identifying positive correlations between germline methylation and broad-scale recombination rates in the human genome, we continued to explore whether DNA methylation is associated with fine-scale recombination patterns. We investigated recombination hotspots, which are usually 1-2kb in length and have significant (usually in orders of magnitude) increase in recombination rate from the background [136, 137]. Based upon the positive correlations between DNA methylation and recombination rate in human sperm, it would be expected that the DNA methylation level of recombination hotspots should be higher than the genomic background. Consistent with this prediction, we found that DNA methylation levels at the species-specific hotspots are significantly higher than the genome average as well as the genomic control regions in the human genome (Figure 3.2A). The genomic control regions have the same CpG number distribution of the recombination hotspots (see Materials and Methods). Interestingly, we found the same pattern in the chimpanzee genome. However, the difference of DNA methylation level between recombination hotspots and genomic background is more significant in human than in chimpanzee (Figure 3.2A).

Several studies have shown that recombination hotspot locations and usage are highly divergent between humans and chimpanzees [140, 141, 151]. Considering that these two species have highly similar genomes, other factors, such as epigenetic factors, may contribute to the evolution of species-specific hotspots. We also utilized the fact that there are small numbers of recombination hotspots that are common between human and chimpanzee genomes. We identified a total of 131 'common' recombination hotspots (see Materials and Methods). The most parsimonious explanation for these common hotspots is that they may have existed in the genome of human and chimpanzee common ancestor: thus species specific hotspots may be evolutionarily 'younger' than the common hotspots. We then compared the variation of recombination rates and DNA methylation levels of 1) common recombination hotspots (n = 131), 2) species-specific

recombination hotspots (n = 9169 and 4906, for humans and chimpanzees, respectively), and 3) the syntenic genomic regions corresponding to the species-specific hotspots of the other species (e.g., human genomic regions syntenic to chimpanzee recombination hotspots while not recombination hotspots in the human genome: n = 4906, chimpanzee genomic regions syntenic to human recombination hotspots while not recombination hotspots in the chimpanzee genomes: n = 9169). The syntenic regions may be considered as genomic 'control' regions for those recently became recombination hotspots in the genome of the other species.

We first examined distributions of recombination rates across these three types of genomic regions. As expected, human-specific recombination hotspots have much higher recombination rates than the syntenic regions of chimpanzee recombination hotspots (Figure 3.2C) and vice versa (Figure 3.2D). We also observed that species-specific recombination hotspots exhibit significantly higher recombination rates than the common recombination hotspots in both species (Figure 3.2C, 3.2D).

We then compared DNA methylation levels of these three types of genomic regions. Interestingly, the patterns observed here do not follow the genome-wide trend of strong correlation between DNA methylation and recombination. In the human genome, syntenic regions of chimpanzee recombination hotspots on average exhibit lower levels of DNA methylation than human-specific recombination hotspots, but significantly higher than common recombination hotspots (Figure 3.2E). In the chimpanzee genome, syntenic regions of human recombination hotspots are significantly more methylated than both the chimpanzee-specific recombination hotspots and common recombination hotspots (Figure 3.2F). Thus, both species exhibit the following pattern of methylation gradient: human recombination hotspots (or regions syntenic to human recombination hotspots) > chimpanzee recombination hotspots (or regions syntenic to chimpanzee recombination hotspots) > common recombination hotspots.

To examine this observation further we performed the following experiment. We calculated inter-species methylation differences between the average fractional methylation level at the human-specific recombination hotspots and their syntenic regions in chimpanzee, as well as the methylation difference between the chimpanzee-specific recombination hotspots and their syntenic regions in human (to facilitate a direct comparison, the difference was always calculated as mean methylation level in human - mean methylation level in chimpanzee). In order to take the methylation difference at the genomic level between human and chimpanzee into account, we calculated the methylation differences by generating the genomic control regions and obtained the distribution of the these methylation differences by bootstrapping one million times (see Materials and Methods). If the correlation between DNA methylation and recombination is consistent at the recombination hotspots with the genome-wide level, we should observe a decreased methylation level difference at the chimpanzee-specific recombination hotspots and increased one at the human-specific recombination hotspots when compared to the genomic control regions. However, methylation differences at both human- and chimpanzee-specific recombination hotspots do not deviate significantly from the distribution of bootstrapped methylation difference (Figure 3.2B). Together, these results demonstrate that DNA methylation may have different effects on fine scale versus broad scale recombination patterns.

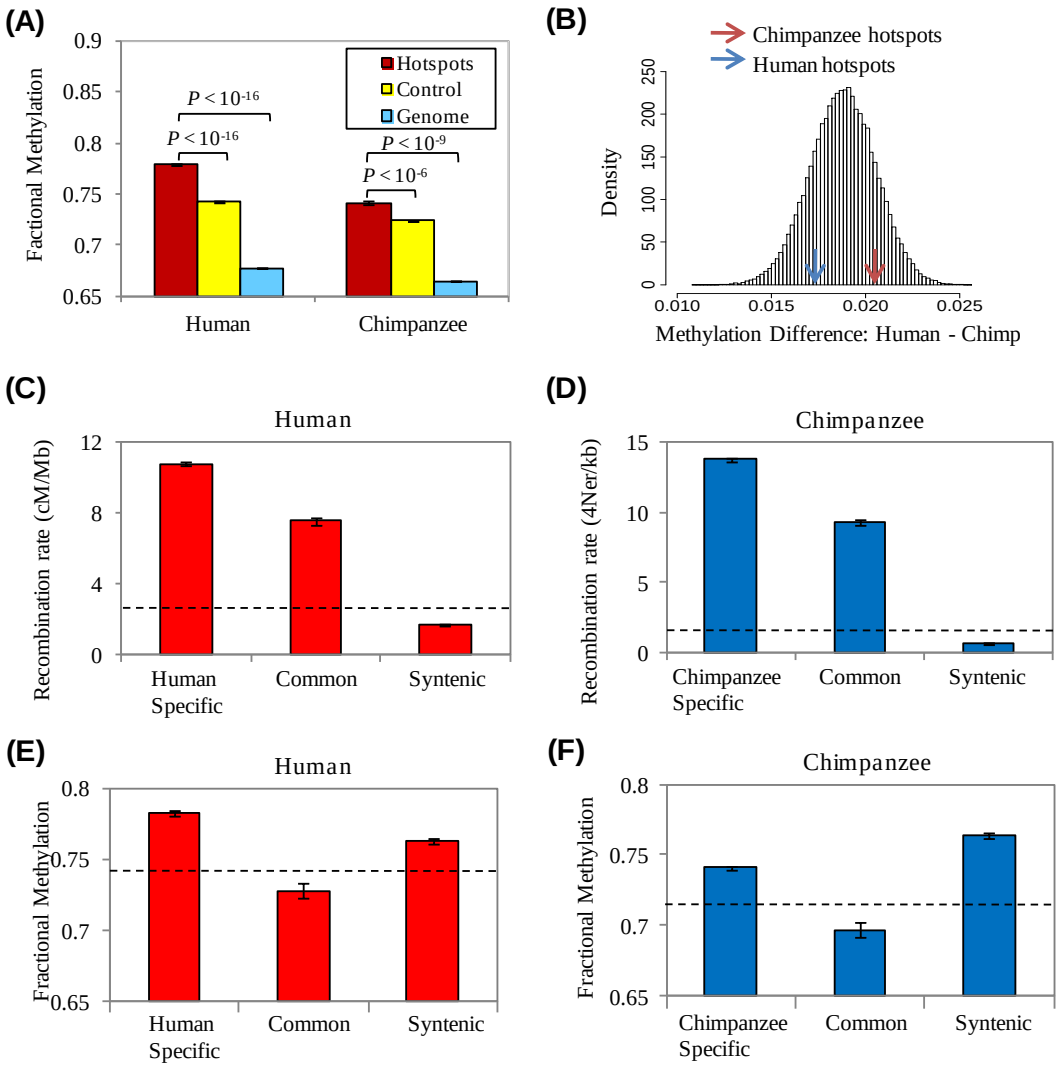

Figure 3.2. Comparison of DNA methylation and recombination rates between recombination hotspots and syntenic regions. (A) Comparison of average fractional DNA methylation level among recombination hotspots, genomic control regions and genome background in human and chimpanzee. (B) Distribution of bootstrapped methylation difference (always calculated as 'Human-Chimp'). The observed inter-species methylation difference are marked for human-specific recombination hotspots (blue arrow) and chimpanzee-specific recombination hotspots (red arrow). Comparison of average recombination rate (C) and fractional DNA methylation level (E) among human-specific recombination hotspots (Human specific), common recombination hotspots (Common) and human syntenic regions of chimpanzee-specific hotspots (Syntenic) in human genome. Comparison of average recombination rate (D) and fractional DNA methylation level (F) among chimpanzee-specific recombination hotspots (Chimpanzee specific), common recombination hotspots (Common) and chimpanzee syntenic regions of human-specific hotspots (Syntenic). For (C), (D), (E) and (F), the average of recombination rate or DNA methylation level for the genomic control regions in the corresponding species are indicated by the dash lines.

Histone modifications and recombination hotspots

The PR domain-containing 9 locus (PRDM9) is an important *trans*-acting factor that controls hotspots specification in both human and mice [154, 161, 162]. This factor can specifically binds to a 13-bp consensus motif that is common to many human hotspots. PRDM9 contains a KR protein-protein binding domain [163], a PR/SET domain that can trimethylate H3K4 [164] and an array of 8 – 16 zinc fingers. It is expressed only during early meiosis, and deficiency of the protein results in abnormal meiosis with aberrant location of DNA strand breaks [164]. In a detailed study of two recombination hotspots in mouse, it was shown that H3K4 trimethylation (H3K4me3) precedes recombination and potentiates hotspot activity [155]. Therefore, we hypothesize that the H3K4me3 profile in germline may affect recombination patterns, especially the location and activity of hotspots. To investigate whether H3K4me3 is a global feature of recombination hotspots, we examined histone modification profiles of human sperms. Human sperm generally lacks histones, as most of histones are replaced with protamines during early germ cell development [159]. Nevertheless, we found that human specific recombination hotspots exhibit over 3-fold enrichment of H3K4me3 enriched regions: we found that 816 human recombination hotspots are overlapped with H3K4me3 enriched regions (see Materials and Methods), while the expected number of overlap is 229 if the H3K4me3 marks were uniformly distributed in the genome (Figure 3.3). This is a highly significant enrichment based upon Fisher's exact test ($P < 10^{-16}$, Figure 3.3). By contrast, neither common recombination hotspots nor syntenic regions to chimpanzee recombination hotspots exhibited statistically significant enrichment (Figure 3.3).

We also examined the distribution of the H3K27me3 mark at the hotspots. We found that the H3K27me3 mark is also significantly over-represented in the human recombination hotspots but not as strongly as the H3K4me3 mark (Figure 3.3). Intriguingly, H3K27me3 is also slightly (1.6-fold) but significantly over-represented at

the human syntenic region of chimpanzee recombination hotspots (Figure 3.3). The number of H3K27me3 enriched regions is also higher than expected in the common recombination hotspots (11 observed compared to 4 expected), but this comparison is not significant due to the small sample size. Moreover, we found that the average fine-scale recombination rates around both the H3K4me3 and H3K27me3 enriched region show an interesting pattern: we found recombination rates are elevated by about 20% and 25% at the H3K4me3 and H3H27me3 enriched regions respectively, when compared to the genomic background (Figure 3.4A).

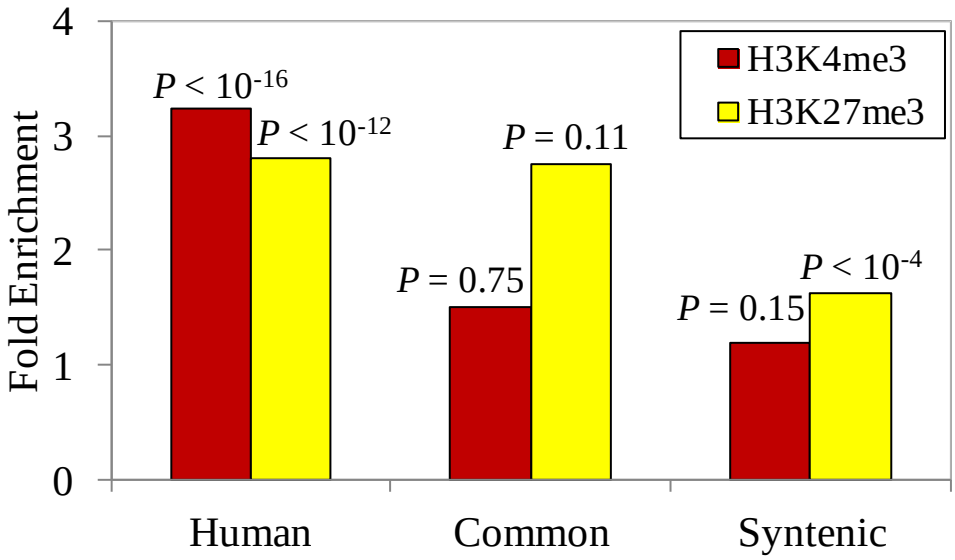

Figure 3.3.Histone modifications are associated with human recombination hotspots. Fold enrichment between observed and expected overlapping of H3K4me3 and H3K27me3 to human-specific recombination hotspots (Human), common recombination hotspots between human and chimpanzee (Common) and human syntenic region of chimpanzee recombination hotspots (Syntenic).

DISCUSSION

Several technical developments, including computational and statistical analyses of single-nucleotide polymorphism (SNP) data, molecular analyses of recombination hotspots in sperm samples and large-scale analyses of pedigrees, are propelling current research of recombination and its mechanistic role in population genetics and

evolutionary processes [165]. Emerging evidence indicates that DNA sequences themselves are not the sole determinants of inter- and intra-species variation in recombination patterns. For example, hotspot locations and usage vary among human individuals, and between humans and chimpanzees [153, 160]. DNA methylation is a strong candidate epigenetic factor that may affect recombination patterns, since it was proven to be established at prophase I in meiosis when recombination occurs [152]. In this study, we thus examined detailed relationships between DNA methylation levels and recombination rates utilizing comprehensive whole genome nucleotide-resolution DNA methylation maps from human and chimpanzee sperms and brains. These nucleotide-resolution DNA methylation maps allow us to investigate fine-scale variation of DNA methylation and correlate with the evolution of recombination hotspots. We found that DNA methylation level is significantly and positively correlated with recombination rate in sperm but not in brain. This indicates that the germline DNA methylation affects variation of broad-scale recombination patterns, while somatic DNA methylation patterns do not.

Sigurdsson et. al [22] have previously explored the co-variation between germline DNA methylation and recombination. Due to the lack of experimentally determined DNA methylation data at that time, they used methylation-associated SNPs (mSNPs) from HapMap data set as a surrogate marker for germline DNA methylation. Even though the result of this study (referred to as 'mSNP' henceforth) also showed a genome-widely positive correlation between mSNPs and regional recombination rate, there are several notable and significant difference between the mSNP study and the current study: 1) mSNPs were generally higher correlated with recombination rate than our study. For example, the correlation coefficient was 0.622 in 500-kb windows from the mSNPs study, while it is 0.212 based on the result from our study. 2) mSNPs was found to be the strongest predictor of recombination rate in a linear model from the ENCODE regions with increased density of known SNPs and sequence information. But our model

indicates that the effect of DNA methylation in recombination rate is weaker than other sequence features (Table 3.2). These two discrepancies can be explained by the idea that mSNP density actually not only reflects DNA methylation levels *per se*, but also other sequence features, such as GC contents, CpG dinucleotide contents and repeat frequencies. Our study thus may provide more realistic representation of the genome-wide relationship between DNA methylation and recombination rates.

Unlike the observation in human genome, we found a slightly but significantly negative correlation between DNA methylation level and recombination rate in the chimpanzee genome. This result is consistent with the finding of recombination elevations at the promoters of genes with high level of DNA methylation in human, but in chimpanzee the elevations occur at the genes with low level of DNA methylation [153]. Therefore, DNA methylation may have species-specific effect in shaping the global and/or regional recombination pattern. Alternatively, it is possible that chimpanzee recombination rates data contains more noise than the human data, and the observed (weak) negative correlation is spurious. Analyses of future more refined chimpanzee genomic recombination rates variation of recombination rates in chimpanzee genomes are necessary to resolve this question.

The third and the most significant difference between our study and that of Sigurdsson et al. is the relationship between fine-scale recombination rates and DNA methylation. In Sigurdsson et al. [22], mSNP frequencies were positively correlated with the number of bases within recombination hot spots in a genome-wide resolution of 125 – 1000 kb, thus the authors claimed that DNA methylation might also affect recombination strongly at fine-scale. However, when we examine experimentally determined DNA methylation levels at human- and chimpanzee-specific recombination hotspots, we find a very intriguing pattern: while recombination rates show significant difference between species-specific recombination hotspots versus syntenic genomic regions corresponding to the recombination hotspots in the other species, DNA

methylation levels do not (Figure 3.2B, E and F). This observation implies that molecular mechanisms linking recombination and DNA methylation may be divergent between fine-scale and broad-scale recombination patterns.

To explore this observation further, we utilized the fact that there are 'common' recombination hotspots shared between human and chimpanzee genomes. Given that recombination hotspots evolve rapidly [141, 142, 160] and that there is no evidence that common recombination hotspots independently evolve in human and chimpanzee genomes, a parsimonious explanation is that these common hotspots represent those that were shared between the two genomes before the evolution of species-specific recombination hotspots. Interestingly, in both species, genomic regions belonging to human recombination hotspots exhibit the highest DNA methylation levels, followed by chimpanzee-specific recombination hotspots, and the common recombination hotspots (Figure 3.2E, 3.2F). These observations suggest that some sequence characteristics can account for the high degree of DNA methylation in both species in spite of the highly divergent inter-species recombination rates. Human recombination hotspots, and chimpanzee genomic regions syntenic to human recombination hotspots, may harbor specific sequence characteristics that are associated with high DNA methylation. Chimpanzee recombination hotspots and human syntenic regions to chimpanzee recombination hotspots also carry some sequence signatures for high levels of DNA methylation. On the other hand common recombination hotspots may have lost some of these sequence features which lead to the decrease of DNA methylation level.

The observed correlation between DNA methylation and recombination rate could due to a third variable, such as histone modification that can interact with both variables and may be more proximal to the cause. The PRDM9 locus plays significant roles generating recombination hotspots [154, 161, 162]. This protein encodes a SET-methyltransferase domain in the *Prdm9* gene, which is responsible for the trimethylation of H3K4 [164]. Mutations in zinc-finger-encoding region of this locus leads to the

change of contact residues in the DNA sequences, provide a simple means of replacing lost hotspots [166]. We thus investigated the association of the H3K4me3 and H3K27me3 to the recombination hotspots using Chip-Seq in human sperm [159]. We and found that both of these histone modifications are significantly enriched at the human-specific recombination hotspots, but not at the common recombination hotspots (Figure 3.3). This result is consistent with the idea that the PRDM9 play a critical role in creating a whole new family of recombination hotspots [166]. This result, together with the observation of elevated recombination rates at the H3K4me3 enriched loci (Figure 3.4A), supports the idea that H3K4me3 may be a global feature at the human recombination hotspots, as it was observed in mouse genome [167]. Our observation of enriched H3K27me3 mark at the human-specific recombination hotspots is partially contradict to the study showing the H3K27me3 is enriched at the *Psmb9* hotspot in the recombinationally inactive mouse strain [155]. However, the previous study only investigated a single recombination hotspot in mouse while our study checked the association of H3K27me3 to the human-specific hotspots genome-widely. In addition, there are 10,621 genomic regions bear both H3K4me3 and H3K27me3 marks, thus termed bivalent regions, and we found that one third of the overlaps between human-specific recombination hotspots and the H3K27me3 mark are from the bivalent regions. These results indicate that the H3K27me3 mark could also be an important molecular feature at the human recombination hotspots and it may affect the recombination pattern simultaneously and interactively with the H3K4me3. This is supported by the observation of ~4 fold-enrichment of the association between recombination hotspots and the bivalent regions ($P < 10^{-16}$).

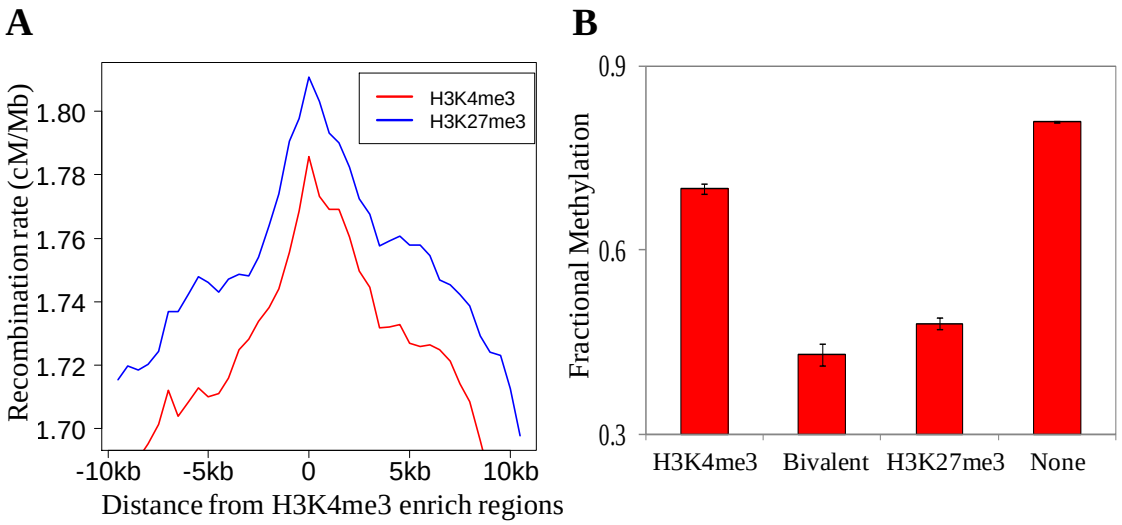

Figure 3.4. The fine-scale profile of recombination rate at histone modifications and their interactions with DNA methylation. (A) Average recombination rate as a function of distance to nearest H3K4me3 and H3K27me3 enrich regions. (B) Comparison of DNA methylation at the recombination hotspots co-localize with H3K4me3 enriched regions alone (H3K4me3), H3K27me3 enriched regions alone (H3K27me3), both H3K4me3 and H3K27me3 enriched regions (Bivalent) and the ones without any overlap of the enriched regions (None).

CHAPTER 4

CONCLUSIONS

As one of the best studied covalent epigenetic modifications, DNA methylation has been investigated in diverse species for the past four decades. From the studies of short individual DNA segments to the high-throughput whole genome analyses, our understanding of the function of DNA methylation, especially its relationship to transcriptional control, is growing fast. Even though some generalizations about the function role of DNA methylation are holding up, new and unexpected phenomena are also being detected all the time, which highlights our limitation in understanding this epigenetic system. The aim of my research was to build an unbiased and comparative framework in order to answer several novel and critical questions regarding the functional role and evolutionary significance of DNA methylation. Toward this end, I first set out to investigate questions regarding how patterns of DNA methylation differ between closely related species and whether such differences contribute to species-specific phenotypes. To investigate these questions, we generated nucleotide-resolution, whole-genome methylation maps of the prefrontal cortex of multiple humans and chimpanzees (methyl-C-seq). This method is a superior choice for comparative studies for a couple of reasons; First, the methyl-C-seq method does not depend on underlying sequences, thus making it ideal to be used in comparisons of genome-wide patterns of DNA methylation between species. Second, because the methyl-C-seq approach enables the methylation frequency of each cytosine to be estimated independently, we can evaluate global differences between methylation maps of different tissues and species. By using this method, we discovered several significant patterns in the brain methylation maps, and inferred potential global-level differences between the brain DNA-methylation maps of humans and chimpanzees. Integrating data on DNA methylation with newly generated data on gene expression, we show that changes in DNA methylation at least

partially explain the divergence of gene-expression patterns in human and chimpanzee brains. Furthermore, differentially methylated genes show striking associations with specific neurological and psychological disorders and cancers, suggesting that changes of DNA methylation might be linked to the evolution of human-specific disease vulnerabilities. In summary, the results of chapter 1 highlight the utility of comparative studies in identifying key epigenomic modifications underlying human-specific phenotypes, including disease vulnerabilities.

In chapter 2, I continued to study the function role of DNA methylation in a comparative frameworks and focused on the human CpG islands, which mark epigenetic regulatory hotspots of mammalian genomes. By performing global analyses of DNA methylation of CpG islands in the human genome, we examined variation of CpG island methylation across multiple methylomes of distinctive cellular origins. This analysis reveals that, contrary to the prevailing notion, CpG islands mark the most highly variably methylated regions in the human genome. Many CpG islands exhibit methylome-specific patterns of DNA methylation. Remarkably, DNA methylation patterns of CpG islands reflect their distinctive nature at many biological levels, including genomic characteristics such as lengths and nucleotide composition, as well as evolutionary features. Moreover, the regulatory functions of CpG islands are tightly linked to their genomic, evolutionary, and DNA methylation features, as evidenced by the co-variation between DNA methylation variability and functional ontology terms and transcriptional profiles. In addition, CpG islands implicated in distinctive biological processes such as diseases, aging, and imprinting exhibit intriguing differences in their genomic, epigenomic and functional features. These new findings from chapter 2 provide novel insights into deciphering the regulatory mechanisms of CpG islands in human health and diseases. What is more important, our results may be used to improve empirical studies of DNA methylation variation across different biological conditions and demography.

Finally, as a way to understand the influence of DNA methylation on primates genome evolution, we investigated the relationship between germline DNA methylation and meiotic recombination, which generates the raw material of evolution and lies at the heart of all genetic analysis. Our genome-wide correlation analyses indicate the positive correlation between DNA methylation and recombination rates is present in germline but not in somatic tissue such as brain. Multiple regression analyses suggest that DNA methylation might be one additional factor affecting recombination in addition to the sequence features. Intriguingly, we observed that DNA methylation has different effect in broad- and fine-scale recombination pattern by comparing both intra- and inter- DNA methylation levels at human- and chimpanzee-specific recombination hotspots. Our results also revealed that DNA methylation may closely interact with histone modifications to simultaneously regulate the fine-scale recombination pattern. The work in chapter 3 sheds lights on the role of epigenetic mechanisms in explaining the phenomenon of inter-individual differences in recombinational activity, despite identical DNA sequence, and also highlights the evolutionary significance of DNA methylation in the human genome

In summary, due to the development of next generation sequencing technique, I got the chance to generate and utilize the whole-genome DNA methylation profile, and thus provide an unbiased and comprehensive view of DNA methylation pattern in human genome as well as in the closely related species, chimpanzee. The three chapters from this dissertations integrate patterns found in genomes, methylomes, and transcriptomes to comprehensively analyze the effect of DNA methylation on the regulation of gene expression and genome evolution. By addressing knowledge gaps and longstanding questions at DNA methylation in human genome, I hope these work can expand our knowledge for this complex epigenetic system, which would finally provide a deeper understanding of the much-needed connections between genotypes and phenotypes.

REFERENCES

1. Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development.* Science, 1975. **187**(4173): p. 226-32.
2. Jones, P.A. and D. Takai, *The role of DNA methylation in mammalian epigenetics.* Science, 2001. **293**(5532): p. 1068-70.
3. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites.* Trends Genet, 1997. **13**(8): p. 335-40.
4. Jones, P.A. and P.W. Laird, *Cancer epigenetics comes of age.* Nature Genetics, 1999. **21**(2): p. 163-7.
5. Robertson, K.D. and A.P. Wolffe, *DNA methylation in health and disease.* Nat Rev Genet, 2000. **1**(1): p. 11-9.
6. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nat Biotechnol, 2010. **28**(10): p. 1057-68.
7. Riggs, A.D., *X inactivation, differentiation, and DNA methylation.* Cytogenet Cell Genet, 1975. **14**(1): p. 9-25.
8. Avner, P. and E. Heard, *X-chromosome inactivation: counting, choice and initiation.* Nat Rev Genet, 2001. **2**(1): p. 59-67.
9. Constancia, M., et al., *Imprinting mechanisms.* Genome Res, 1998. **8**(9): p. 881-900.
10. Wagschal, A. and R. Feil, *Genomic imprinting in the placenta.* Cytogenet Genome Res, 2006. **113**(1-4): p. 90-8.
11. Bestor, T.H., *The DNA methyltransferases of mammals.* Hum Mol Genet, 2000. **9**(16): p. 2395-402.
12. Maunakea, A.K., et al., *Conserved role of intragenic DNA methylation in regulating alternative promoters.* Nature, 2010. **466**(7303): p. 253-7.
13. Shukla, S., et al., *CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.* Nature, 2011. **479**(7371): p. 74-U99.
14. Choy, J.S., et al., *DNA methylation increases nucleosome compaction and rigidity.* J Am Chem Soc, 2010. **132**(6): p. 1782-3.
15. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, 1992. **69**(6): p. 915-26.
16. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.
17. Weber, M., et al., *Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.* Nature Genetics, 2005. **37**(8): p. 853-62.
18. Hellman, A. and A. Chess, *Gene body-specific methylation on the active X chromosome.* Science, 2007. **315**(5815): p. 1141-3.
19. Suzuki, M.M., et al., *CpG methylation is targeted to transcription units in an invertebrate genome.* Genome Res, 2007. **17**(5): p. 625-31.
20. Feng, S., et al., *Conservation and divergence of methylation patterning in plants and animals.* Proc Natl Acad Sci U S A, 2010. **107**(19): p. 8689-94.
21. Zemach, A., et al., *Genome-wide evolutionary analysis of eukaryotic DNA methylation.* Science, 2010. **328**(5980): p. 916-9.

22. Sigurdsson, M.I., et al., *HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination.* Genome Res, 2009. **19**(4): p. 581-9.
23. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
24. Karaman, M.W., et al., *Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts.* Genome Res, 2003. **13**(7): p. 1619-30.
25. Khaitovich, P., et al., *Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.* Science, 2005. **309**(5742): p. 1850-4.
26. King, M.-C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-116.
27. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.
28. Feil, R., *Environmental and nutritional effects on the epigenetic regulation of genes.* Mutat Res, 2006. **600**(1-2): p. 46-57.
29. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
30. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
31. *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.
32. Backes, C., et al., *GeneTrail--advanced gene set enrichment analysis.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W186-92.
33. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
34. Homer, N., B. Merriman, and S.F. Nelson, *BFAST: an alignment tool for large scale genome resequencing.* PLoS One, 2009. **4**(11): p. e7767.
35. Shi, L., et al., *The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.* Nat Biotechnol, 2010. **28**(8): p. 827-38.
36. Miller, J.A., M.C. Oldham, and D.H. Geschwind, *A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging.* J Neurosci, 2008. **28**(6): p. 1410-20.
37. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation.* Genome Res, 2010. **20**(3): p. 320-31.
38. Li, Y., et al., *The DNA methylome of human peripheral blood mononuclear cells.* PLoS Biol, 2010. **8**(11): p. e1000533.
39. Enard, W., et al., *Differences in DNA methylation patterns between humans and chimpanzees.* Current Biology, 2004. **14**(4): p. R148-9.
40. Gama-Sosa, M.A., et al., *Tissue-specific differences in DNA methylation in various mammals.* Biochim Biophys Acta, 1983. **740**(2): p. 212-9.
41. Elango, N. and S.V. Yi, *DNA methylation and structural and functional bimodality of vertebrate promoters.* Mol Biol Evol, 2008. **25**(8): p. 1602-8.

42. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nature Genetics, 2007. **39**(4): p. 457-66.
43. Suzuki, M.M. and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics.* Nat Rev Genet, 2008. **9**(6): p. 465-76.
44. Benjamini, Y. and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency.* Annals of Statistics, 2001. **29**(4): p. 1165-1188.
45. Storey, J.D., *A direct approach to false discovery rates.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2002. **64**: p. 479-498.
46. Aran, D., et al., *Replication timing-related and gene body-specific methylation of active human genes.* Hum Mol Genet, 2011. **20**(4): p. 670-80.
47. Gilad, Y., et al., *Expression profiling in primates reveals a rapid evolution of human transcription factors.* Nature, 2006. **440**(7081): p. 242-5.
48. Agis-Balboa, R.C., et al., *A hippocampal insulin-growth factor 2 pathway regulates the extinction of fear memories.* Embo Journal, 2011. **30**(19): p. 4071-83.
49. Hawkins, N.A., et al., *Neuronal voltage-gated ion channels are genetic modifiers of generalized epilepsy with febrile seizures plus.* Neurobiol Dis, 2011. **41**(3): p. 655-60.
50. Trudeau, M.M., et al., *Heterozygosity for a protein truncation mutation of sodium channel SCN8A in a patient with cerebellar atrophy, ataxia, and mental retardation.* J Med Genet, 2006. **43**(6): p. 527-30.
51. Han, L., et al., *DNA methylation regulates MicroRNA expression.* Cancer Biol Ther, 2007. **6**(8): p. 1284-8.
52. Pai, A.A., et al., *A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues.* PLoS Genet, 2011. **7**(2): p. e1001316.
53. Brawand, D., et al., *The evolution of gene expression levels in mammalian organs.* Nature, 2011. **478**(7369): p. 343-8.
54. Huh, I., et al., *DNA methylation and transcriptional noise.* Epigenetics Chromatin, 2013. **6**(1): p. 9.
55. Numata, S., et al., *DNA methylation signatures in development and aging of the human prefrontal cortex.* Am J Hum Genet, 2012. **90**(2): p. 260-72.
56. Siegmund, K.D., et al., *DNA methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons.* PLoS One, 2007. **2**(9): p. e895.
57. Bocker, M.T., et al., *Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging.* Blood, 2011. **117**(19): p. e182-9.
58. Christensen, B.C., et al., *Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context.* PLoS Genet, 2009. **5**(8): p. e1000602.
59. Hernandez, D.G., et al., *Distinct DNA methylation changes highly correlated with chronological age in the human brain.* Hum Mol Genet, 2011. **20**(6): p. 1164-72.
60. Caceres, M., et al., *Elevated gene expression levels distinguish human from non-human primate brains.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 13030-5.

61. Gu, J. and X. Gu, *Induced gene expression in human brain after the split from chimpanzee.* Trends Genet, 2003. **19**(2): p. 63-5.
62. Preuss, T.M., et al., *Human brain evolution: insights from microarrays.* Nat Rev Genet, 2004. **5**(11): p. 850-60.
63. Olson, M.V. and A. Varki, *Sequencing the chimpanzee genome: insights into human evolution and disease.* Nat Rev Genet, 2003. **4**(1): p. 20-8.
64. Li, S., et al., *Environmental exposure, DNA methylation, and gene regulation: lessons from diethylstilbesterol-induced cancers.* Ann N Y Acad Sci, 2003. **983**: p. 161-9.
65. Weaver, I.C., et al., *Epigenetic programming by maternal behavior.* Nat Neurosci, 2004. **7**(8): p. 847-54.
66. Suzuki, M.M. and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics.* Nat. Rev. Genet., 2008. **9**: p. 465-476.
67. Bird, A., *CpG-rich islands and the function of DNA methylation.* Nature, 1986. **321**: p. 209-213.
68. Bird, A., et al., *A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA.* Cell, 1985. **40**(1): p. 91-9.
69. Cooper, D.N., M.H. Taggart, and A.P. Bird, *Unmethylated domains in vertebrate DNA.* Nuc. Acids Research, 1983. **11**(3): p. 647-658.
70. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J. Mol. Biol., 1987. **196**(2): p. 261-282.
71. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proc. Natl. Acad. Sci. USA, 2002. **99**(6): p. 3740-3745.
72. Elango, N. and S.V. Yi, *DNA methylation and structural and functional bimodality of vertebrate promoters.* Mol. Biol. Evol., 2008. **25**: p. 1602-1608.
73. Larsen, F., et al., *CpG islands as gene markers in the human genome.* Genomics, 1992. **13**(4): p. 1095-1107.
74. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nat. Genet., 2007. **39**(4): p. 457-466.
75. Guenther, M.G., et al., *A chromatin landmark and transcription initiation at most promoters in human cells.* Cell, 2007. **130**(1): p. 77-88.
76. Vavouri, T. and B. Lehner, *Human genes with CpG island promoters have a distinct transcription-associated chromatin organization.* Genome Biology, 2012. **13**(11): p. R110.
77. Carninci, P., *Tagging mammalian transcription complexity.* Trends Genet., 2006. **22**(9): p. 501-510.
78. Yaragatti, M., C. Basilico, and L. Dailey, *Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions.* Genome Res., 2008. **18**(6): p. 930-938.
79. Williams, S., et al., *CpG-island fragments from the HNRPA2B1/CBX3 genomic locus reduce silencing and enhance transgene expression from the hCMV promoter/enhancer in mammalian cells.* BMC Biotechnology, 2005. **5**(1): p. 17.
80. Xi, H., et al., *Identification and Characterization of Cell Type-Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome.* PLoS Genetics, 2007. **3**(8): p. e136.

81. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nat Biotech, 2010. **28**(10): p. 1057-1068.
82. Robertson, K.D. and A.P. Wolffe, *DNA methylation in health and disease.* Nat. Rev. Genet., 2000. **1**: p. 11-19.
83. Elango, N. and S.V. Yi, *Functional relevance of CpG island length for regulation of gene expression.* Genetics, 2011. **187**: p. 1077-1083.
84. Fenouil, R., et al., *CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters.* Genome Research, 2012. **in press**.
85. Cohen, N.M., E. Kenigsberg, and A. Tanay, *Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection.* Cell, 2011. **145**(5): p. 773-786.
86. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation.* Genome Research, 2010. **20**(3): p. 320-331.
87. Li, Y., et al., *The DNA methylome of human peripheral blood mononuclear cells.* PLoS Biol, 2010. **8**(11): p. e1000533.
88. Zeng, J., et al., *Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution.* The American Journal of Human Genetics, 2012. **91**(3): p. 455-465.
89. Molaro, A., et al., *Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates.* Cell, 2011. **146**(6): p. 1029-1041.
90. Schroeder, D.I., et al., *The human placenta methylome.* Proceedings of the National Academy of Sciences, 2013.
91. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucl. Acids Res., 2008. **36**(suppl_1): p. D773-779.
92. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-322.
93. Brawand, D., et al., *The evolution of gene expression levels in mammalian organs.* Nature, 2011. **478**(7369): p. 343-348.
94. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc. Nat. Acad. Sci. USA, 2004. **101**(16): p. 6062-6067.
95. Yanai, I., et al., *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.* Bioinformatics, 2005. **21**(5): p. 650-659.
96. Lister, R., et al., *Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.* Nature, 2011. **471**(7336): p. 68-73.
97. Imamura, T., et al., *CpG island of rat sphingosine kinase-1 gene: tissue-dependent DNA methylation status and multiple alternative first exons.* Genomics, 2001. **76**(1-3): p. 117-25.
98. Futscher, B.W., et al., *Role for DNA methylation in the control of cell type specific maspin expression.* Nature Genetics, 2002. **31**(2): p. 175-9.
99. Shiota, K., et al., *Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice.* Genes Cells, 2002. **7**(9): p. 961-9.
100. Schilling, E. and M. Rehli, *Global, comparative analysis of tissue-specific promoter CpG methylation.* Genomics, 2007. **90**(3): p. 314-23.

101. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.
102. Ji, H., et al., *Comprehensive methylome map of lineage commitment from haematopoietic progenitors.* Nature, 2010. **467**(7313): p. 338-342.
103. Baron, U., et al., *DNA demethylation in the human FOXP3 locus discriminates regulatory T cells from activated FOXP3+ conventional T cells.* European Journal of Immunology, 2007. **37**(9): p. 2378-2389.
104. Newman, J.R.S., et al., *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise.* Nature, 2006. **441**: p. 840-846.
105. Yin, S., et al., *Dosage compensation on the active X chromosome minimizes transcriptional noise of X-linked genes in mammals.* Genome Biology, 2009. **10**(7): p. R74.
106. Huh, I., et al., *DNA methylation and transcriptional noise.* Epigenetics & Chromatin, 2013. **6**(1): p. 9.
107. Sanford, J.P., et al., *Differences in DNA methylation during oogenesis and spermatogenesis and their persistence during early embryogenesis in the mouse.* Genes & Development, 1987. **1**(10): p. 1039-1046.
108. De Smet, C., et al., *DNA Methylation Is the Primary Silencing Mechanism for a Set of Germ Line- and Tumor-Specific Genes with a CpG-Rich Promoter.* Molecular and Cellular Biology, 1999. **19**(11): p. 7327-7335.
109. Farthing, C.R., et al., *Global Mapping of DNA Methylation in Mouse Promoters Reveals Epigenetic Reprogramming of Pluripotency Genes.* PLoS Genet, 2008. **4**(6): p. e1000116.
110. ENCODE_consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
111. Kim, S.-H., et al., *Heterogeneous genomic molecular clocks in primates.* PLoS Genetics, 2006. **2**: p. e163.
112. Elango, N., et al., *Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation.* PLoS Computational Biology, 2008. **4**(2): p. e1000015.
113. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes & Development, 2011. **25**(10): p. 1010-1022.
114. Landolin, J.M., et al., *Sequence features that drive human promoter function and tissue specificity.* Genome Research, 2010. **20**: p. 890-898.
115. Sproul, D., et al., *Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns.* Genome Biology, 2012. **13**(10): p. R84.
116. Heyn, H., et al., *Distinct DNA methylomes of newborns and centenarians.* Proceedings of the National Academy of Sciences, 2012. **109**(26): p. 10522-10527.
117. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-492.
118. Bernstein, B.E., et al., *Genomic maps and comparative analysis of histone modifications in human and mouse.* Cell, 2005. **120**: p. 169-181.
119. Eckhardt, F., et al., *DNA methylation profiling of human chromosomes 6, 20 and 22.* Nat Genet, 2006. **38**(12): p. 1378-1385.

120. Ryba, T., et al., *Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types.* Genome Research, 2010. **20**(6): p. 761-770.
121. Kobayashi, H., et al., *Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks.* PLoS Genet, 2012. **8**(1): p. e1002440.
122. Houseman, E., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC Bioinformatics, 2012. **13**(1): p. 86.
123. Liu, Y., et al., *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.* Nat Biotech, 2013. **31**(2): p. 142-147.
124. Fraser, H., et al., *Population-specificity of human DNA methylation.* Genome Biology, 2012. **13**(2): p. R8.
125. Koestler, D.C., et al., *Peripheral Blood Immune Cell Methylation Profiles Are Associated with Nonhematopoietic Cancers.* Cancer Epidemiology Biomarkers & Prevention, 2012.
126. Barzel, A. and M. Kupiec, *Finding a match: how do homologous sequences get together for recombination?* Nature Reviews Genetics, 2008. **9**(1): p. 27-37.
127. Nachman, M.W., *Variation in recombination rate across the genome: evidence and implications.* Curr Opin Genet Dev, 2002. **12**(6): p. 657-63.
128. Duret, L. and P.F. Arndt, *The impact of recombination on nucleotide substitutions in the human genome.* PLoS Genet, 2008. **4**(5): p. e1000071.
129. Charlesworth, B., P. Sniegowski, and W. Stephan, *The evolutionary dynamics of repetitive DNA in eukaryotes.* Nature, 1994. **371**(6494): p. 215-20.
130. Comeron, J.M., M. Kreitman, and M. Aguade, *Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila.* Genetics, 1999. **151**(1): p. 239-49.
131. Coop, G., et al., *High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans.* Science, 2008. **319**(5868): p. 1395-8.
132. Kong, A., et al., *A high-resolution recombination map of the human genome.* Nature Genetics, 2002. **31**(3): p. 241-7.
133. Paigen, K., et al., *The recombinational anatomy of a mouse chromosome.* PLoS Genet, 2008. **4**(7): p. e1000119.
134. Jeffreys, A.J., A. Ritchie, and R. Neumann, *High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot.* Hum Mol Genet, 2000. **9**(5): p. 725-33.
135. Ng, S.H., et al., *A quantitative assay for crossover and noncrossover molecular events at individual recombination hotspots in both male and female gametes.* Genomics, 2008. **92**(4): p. 204-9.
136. Steinmetz, M., et al., *A molecular map of the immune response region from the major histocompatibility complex of the mouse.* Nature, 1982. **300**(5887): p. 35-42.
137. Jeffreys, A.J., L. Kauppi, and R. Neumann, *Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex.* Nature Genetics, 2001. **29**(2): p. 217-22.

138. Arnheim, N., P. Calabrese, and M. Nordborg, *Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved.* American journal of human genetics, 2003. **73**(1): p. 5-16.
139. Myers, S., et al., *The distribution and causes of meiotic recombination in the human genome.* Biochem Soc Trans, 2006. **34**(Pt 4): p. 526-30.
140. Winckler, W., et al., *Comparison of fine-scale recombination rates in humans and chimpanzees.* Science, 2005. **308**(5718): p. 107-11.
141. Ptak, S.E., et al., *Fine-scale recombination patterns differ between chimpanzees and humans.* Nature Genetics, 2005. **37**(4): p. 429-34.
142. Yi, S. and W.H. Li, *Molecular evolution of recombination hotspots and highly recombining pseudoautosomal regions in hominoids.* Molecular biology and evolution, 2005. **22**(5): p. 1223-30.
143. Jeffreys, A.J. and R. Neumann, *The rise and fall of a human recombination hot spot.* Nature Genetics, 2009. **41**(5): p. 625-9.
144. Chinnici, J.P., *Modification of recombination frequency in Drosophila. I. Selection for increased and decreased crossing over.* Genetics, 1971. **69**(1): p. 71-83.
145. Charlesworth, B., *Mutation-selection balance and the evolutionary advantage of sex and recombination.* Genet Res, 2007. **89**(5-6): p. 451-73.
146. Wilfert, L., J. Gadau, and P. Schmid-Hempel, *Variation in genomic recombination rates among animal taxa and the case of social insects.* Heredity (Edinb), 2007. **98**(4): p. 189-97.
147. Groenen, M.A., et al., *A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate.* Genome Research, 2009. **19**(3): p. 510-519.
148. Meunier, J. and L. Duret, *Recombination drives the evolution of GC-content in the human genome.* Mol Biol Evol, 2004. **21**(6): p. 984-90.
149. Marais, G., *Biased gene conversion: implications for genome and sex evolution.* Trends Genet, 2003. **19**(6): p. 330-8.
150. Jensen-Seaman, M.I., et al., *Comparative recombination rates in the rat, mouse, and human genomes.* Genome Res, 2004. **14**(4): p. 528-38.
151. Neumann, R. and A.J. Jeffreys, *Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation.* Hum Mol Genet, 2006. **15**(9): p. 1401-11.
152. Oakes, C.C., et al., *Developmental acquisition of genome-wide DNA methylation occurs prior to meiosis in male germ cells.* Dev Biol, 2007. **307**(2): p. 368-79.
153. Auton, A., et al., *A fine-scale chimpanzee genetic map from population sequencing.* Science, 2012. **336**(6078): p. 193-8.
154. Baudat, F., et al., *PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice.* Science, 2010. **327**(5967): p. 836-840.
155. Buard, J., et al., *Distinct histone modifications define initiation and repair of meiotic recombination in the mouse.* Embo Journal, 2009. **28**(17): p. 2616-24.
156. Borde, V., et al., *Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites.* Embo Journal, 2009. **28**(2): p. 99-111.

157. Zeng, J., et al., *Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution.* Am J Hum Genet, 2012. **91**(3): p. 455-65.
158. Molaro, A., et al., *Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates.* Cell, 2011. **146**(6): p. 1029-41.
159. Hammoud, S.S., et al., *Distinctive chromatin in human sperm packages genes for embryo development.* Nature, 2009. **460**(7254): p. 473-8.
160. Myers, S., et al., *A fine-scale map of recombination rates and hotspots across the human genome.* Science, 2005. **310**(5746): p. 321-4.
161. Berg, I.L., et al., *PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans.* Nature Genetics, 2010. **42**(10): p. 859-63.
162. Myers, S., et al., *Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination.* Science, 2010. **327**(5967): p. 876-9.
163. Birtle, Z. and C.P. Ponting, *Meisetz and the birth of the KRAB motif.* Bioinformatics, 2006. **22**(23): p. 2841-5.
164. Hayashi, K., K. Yoshida, and Y. Matsui, *A histone H3 methyltransferase controls epigenetic events required for meiotic prophase.* Nature, 2005. **438**(7066): p. 374-8.
165. Paigen, K. and P. Petkov, *Mammalian recombination hot spots: properties, control and evolution.* Nat Rev Genet, 2010. **11**(3): p. 221-33.
166. Oliver, P.L., et al., *Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa.* PLoS Genet, 2009. **5**(12): p. e1000753.
167. Smagulova, F., et al., *Genome-wide analysis reveals novel molecular features of mouse recombination hotspots.* Nature, 2011. **472**(7343): p. 375-8.