Distribution Agreement

Signature:

_______________________________ _______________

Elizabeth Marie Kennedy Date

Epigenome-Wide Patterns of DNA Methylation in Radiation Exposure
and Gene Expression

By

Elizabeth Marie Kennedy
Doctor of Philosophy
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

____
Karen N. Conneely, Ph.D.
Advisor

____
Victor G. Corces, Ph.D.
Committee Member

____
Michael P. Epstein, Ph.D.
Committee Member

____
Alicia K. Smith, Ph.D.
Committee Member

____
Paula M. Vertino, Ph.D.
Committee Member

Accepted:

____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____
Date

Epigenome-Wide Patterns of DNA Methylation in Radiation Exposure
and Gene Expression

By

Elizabeth Marie Kennedy
M.S., University of West Florida, 2011

Advisor: Karen N. Conneely, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2017

Abstract

Epigenome-Wide Patterns of DNA Methylation in Radiation Exposure and Gene Expression

By Elizabeth Marie Kennedy

DNA methylation is the most fundamental example of an epigenetic modification and is an integral epigenetic mechanism in humans. Through pathways that are not fully elucidated, DNA methylation can modulate gene transcription, and its patterns change readily over time in response to environmental or stochastic factors. For example, nearly identical methylation patterns among twins diverge over time, in a process known as epigenetic drift. Two natural questions that arise from this information are: how do DNA methylation patterns change in response to environment, and what are the downstream effects of those changes? Through my dissertation work, I have attempted to address both of these questions. First, I present a thorough review of extant literature in the epigenomics of radiation exposure. Second, I present a study that addresses acute and long-term changes to genome-wide CpG methylation patterns that occur following irradiation with varying qualities and quantities of radiation. We found that iron-ion, silicon-ion and X-ray irradiation induced rapid and stable changes in DNA methylation at distinct subsets of CpG sites. Importantly, we found that iron-irradiation-associated CpG sites could differentiate tumor and normal tissues for two human lung cancers. This study suggests that environmental exposures, like radiation, leave a lasting epigenetic imprint, and that these sites may be relevant to the development of complex diseases. Lastly, I present work that aimed to characterize and explore how DNA methylation patterns interact with gene expression, throughout the genome. Among CpGs at which methylation significantly associated with transcription (eCpGs), <50% fell within the canonical promoter region of the associated gene. Rather, we found that eCpGs were more common within enhancer and insulator elements and non-coding RNAs. We suggest that most changes in DNA methylation correlate negatively with transcription, and contrast our findings with the research that established opposing conventional wisdom. My dissertation work sheds new light on the interplay of the epigenome with the environment and with gene expression. Further, this work provides vital and biologically-relevant context for the interpretation of many existing and future studies of DNA methylation.

Epigenome-Wide Patterns of DNA Methylation in Radiation Exposure and Gene Expression

By

Elizabeth Marie Kennedy
M.S., University of West Florida, 2011

Advisor: Karen N. Conneely, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2017

Acknowledgments

I would like to acknowledge my family and friends for the love and support that allowed me to follow my dreams; my children for teaching me about priorities and the value of taking a break; and my advisor and mentor, Karen, who was always exactly the mentor I needed.

Table of Contents

Table of Tables

Table of Supplementary Tables

Table of Figures

Table of Supplemental Figures

CHAPTER I. Introduction

DNA methylation is the cornerstone of epigenetics

Epigenetics is the study of modifications to the genome that affect gene activity and expression, but do not change the underlying sequence of DNA. The most fundamental and well-known example of epigenetic modification is DNA methylation. DNA methylation is the addition of a methyl- group to the carbon 5 position on the deoxyribonucleotide, cytosine (5-methylcytsosine or 5mC). Across different organisms, this chemical modification occurs in three genomic contexts: at adjacent cytosine and guanine dinucleotides (CpG), CHG trinucleotides and CHH trinucleotides (where H is an adenine, thymine, or cytosine). CpG methylation, however, seems to be the primary context in the human genome among fully differentiated cells [1].

CpG methylation is established and maintained through cell generations by the action of DNA methyltransferases. DNA methyltransferases are a group of enzymes that facilitate the transfer of a methyl group from the methyl donor, S-adenosyl methionine, to the DNA. In humans, the three known DNA methyltransferases are DNMT1, DNMT3A and DNMT3B. The most abundant methyltransferase in human cells is DNMT1, which methylates cytosines opposite a CpG on the other strand. Although DNMT1 can catalyze *de novo* methylation, it is primarily known as a maintenance enzyme, as it has higher catalytic activity for the hemimethylated DNA that results from the synthesis of a new DNA strand in the cell cycle. DNMT3A and DNMT3B also catalyze DNA methylation in the CpG context, but have no preference for hemimethylated DNA and so are primarily thought of as *de novo* methyltransferases (reviewed in [2]).

Approximately 70–80% of CpGs in the human genome are methylated and of those, more than 90% lie within the quiescent parts of the genome composed of tightly-packed DNA, called heterochromatin [1, 3, 4]. Heterochromatin tends to be depleted of CpGs, because methylated cytosines are prone to spontaneous deamination, which results in the transition mutation from cytosine to thymine. The transition results in a T:G mismatch that, if not corrected, will be complemented with adenosine during the cell cycle. Over time, this process has led to the depletion of CpGs in areas of the genome that are typically methylated [4]. Unmethylated CpGs are generally found in GC-rich sequences termed CpG islands (CGI) that constitute promoter regions and transcription factor binding sites for more than half of mammalian genes, including nearly all housekeeping genes [3]. Because CpG islands are rarely methylated, they are precluded from deamination and have higher CpG frequencies than the rest of the genome [4].

DNA methylation is an integral epigenetic mechanism in humans. DNA methylation can modulate gene transcription by physically inhibiting the binding of transcription factors to a gene's promoter, as most commonly seen in stable repressed-state genes (imprinted genes, X-inactivation and germline maintenance genes). DNA methylation may also recruit enzymes that modify the proteins that package DNA into structural units, called histones. By modifying the N-terminal tail of histone proteins, histone-modifying enzymes can increase or decrease the transcriptional potential of DNA. Through its modulation of gene expression, DNA methylation (along with other epigenetic modifications, like to histones), allows for morphologically distinct cell types to form from a single genome [2, 4].

DNA methylation is a dynamic mark in the genome

Unlike DNA sequence, DNA methylation patterns change readily over time in response to intrinsic (i.e. genetic) or extrinsic (environmental or stochastic) factors. Methylation patterns at many CpG sites are more similar among related individuals than unrelated individuals, even when those families that do not share a living environment [5, 6]. Additionally, many monozygotic twins acquire methylation at shared loci [7]. These findings suggest a genetic component in methylation patterning. However, at other CpGs, nearly identical methylation patterns among newborn monozygotic twins diverge over time. This difference is reflected in significantly different methylation and histone acetylation patterns later in life, with the greatest epigenetic differences being present in twins that spent more time apart [8, 9]. This age-related divergence in the methylation patterns of twins suggests that stochastic and environmental processes play a large role in DNA methylation changes. Such changes are frequently referred to as epigenetic drift [10].

Radiation exposure as a model for environmental exposures. On Earth, we are exposed to ionizing radiation through diagnostic and therapeutic medical devices, background radiation, cosmic rays, radioactive waste, radon decay, nuclear tests, and nuclear accidents [11]. Astronauts are exposed to higher levels of radiation during space travel. For the most damaging forms of Galactic Cosmic Radiation (GCR), there are not currently effective means of shielding [12]. As space tourism and interplanetary travel become an impending reality, rather than a flight of fancy, it is imperative that we consider how environmental exposures affect the epigenome, as well as the genome [13].

Ionizing radiation exists in different forms. First, there is photon radiation, like x- or γ-radiation, which has a lower linear energy transfer (LET) per dose of radiation; meaning that when it encounters matter (like human cells), it is sparsely ionizing [14]. Second, there is particle radiation. Particle radiation is made up of elements that have been stripped on their electron shells. Simple protons, and alpha particles (Hydrogen and Helium nuclei, respectively), along with simple electrons (beta particles), make up the most common types of GCR, which along with x- or γ-radiation, make up the most common types of terrestrial radiation [15, 16]. Particle radiation heavier than helium is termed HZE radiation, which denotes its high (H) atomic number (Z) and energy (E). HZE particles have a very high LET, meaning that they are densely ionizing [17]. Additionally, HZE radiation generates electrostatic secondary radiation, called δ-radiation, that extends laterally from its main track, increasing the area affected by radiation [18, 19].

Ionizing radiation can cause damage to the genome in a number of ways. Direct damage is caused by a direct interaction of photon or particle radiation with the DNA strand. Indirect damage is caused by the reactive oxygen species (ROS). ROS include the highly reactive compounds, superoxide, hydroxyl radicals, and hydrogen peroxide (H_2O_2). Once generated, ROS will interact with the surrounding organic material including DNA, which can result in oxidative damage that contributes to lasting mutation and carcinogenesis [20]. Whereas low-LET radiation types will interact sparsely with cellular components, resulting primarily in oxidative damage and some single and double stranded DNA breaks, high-LET radiation ionizes densely along the particle track, resulting in complex DNA breakages and high ROS production [11, 15]. Exposure to both high- and low-LET radiation can cause changes to the epigenome.

Both double stranded breaks and oxidative DNA damage cause transient changes to chromatin as part of the cellular DNA damage response. Through a combination of chromatin-level modifications, including repressive histone modifications (like methylation of histone H3 at lysine 27 and histone deacetylation) and DNA methylation by DNMTs, the area around the DNA damage undergoes transcriptional inhibition, to allow for the repair of DNA damage. Although the repair process typically resolves with the resetting of the original chromatin environment, in environments of chronic stress (like the continuous exposure to radiation experienced in space), these transient epigenetic changes have the potential to become permanent ([21], reviewed in [11]). The epigenetic responses to DNA damage are particularly relevant to human health because they are similar to the epigenetic changes observed in cancer cells [22].

Functional role of DNA methylation in the genome

DNA methylation blocks transcription factor binding. DNA methylation has the potential to regulate gene expression [2]. The effects of methylation are commonly studied at gene promoters. CpG islands found at the promoters of many housekeeping or developmentally regulated genes are constitutively hypomethylated. In these promoters, the presence of methyl groups from 5mC in the major groove of the DNA double helix are thought to exclude the binding of transcription factors that drive gene expression. However, the binding of transcription factors at promoters may also be essential to precluding DNMTs that initiate de novo methylation [2, 23].

DNA methylation promotes higher-level epigenetic transcriptional repression. In addition to directly inhibiting the binding of transcription factors, methylated CpGs at CGI also repress gene expression by attracting repressive chromatin modifications. Members of the methyl-CpG binding domain (MBD) family, MBD1, MBD2, and MeCP2, are implicated in methylation-dependent transcriptional repression through association with different co-repressor complexes [24]. For example, MBD1 associates with the histone 3 lysine 9 methyl- transferase SETDB1, ensuring stable silencing of MBD1-associated (and thus, 5mC-associated) genes [25]. Additionally, MBD2 and the NuRD co-repressor complex work in concert as part of the large multi-protein complex, MeCP1 [26]. The NuRD complex is comprised, in part, of histone deacetylases (HDAC), which inhibit transcription through the removal of transcriptional activation marks [27]. In fact, most MBD family proteins associate with HDACs, leading to the inverse correlation of DNA methylation and histone acetylation [24].

Interactions of DNA methylation and histone modifications. The exact relationship of all epigenetic modifications and transcriptional potential are not always clear and sometimes context specific. While the examples above suggest that DNA methylation informs higher-level epigenetic modifications, there are times when the inverse appears to be true. For example, trimethylation of histone 3 lysine 4 (H3K4me3) is a key feature of active gene promoters [28]. Setd1, an H3K4 methyltransferase, is recruited to CGIs through its interaction with the CpG binding protein Cfp1, which binds only unmethylated CGIs [29]. The enrichment of H3K4me3 in gene promoters is thought to prevent *de novo* methylation at CGI. In evidence of this finding, Dnmt3L, a catalytically inactive DNMT that guides the action of *de novo* DNMTs Dnmt3a and Dnmt3b, selectively recognizes nucleosomes that lack H3K4 methylation [2].

Non-promoter targets of epigenetic transcriptional regulation. Although the role of CpG methylation at promoters has been intensively scrutinized, recent research indicates that gene expression can be affected distally by DNA methylation at enhancers and insulators, as well as within the gene itself [4, 30–41]. Further, non-coding RNAs (ncRNAs) function as important regulators of gene expression [34, 37, 40], have been associated with complex diseases and cancers [34, 42], and are sensitive to DNA methylation [43]. The roles of DNA methylation in these genomic features are active areas of debate and research.

Enhancers. Enhancers, as distal regulatory elements, promote gene expression [4, 31, 44]. Enhancers are typically several hundred base pairs long and contain clusters of recognition sequences for transcriptional co-activators or transcription factors [45]. Unlike promoter sequences, enhancers function independently of orientation and position relative to their target gene [46, 47]. They may function from distances ranging from a few kb to 1 Mb from their target. In human CD4$^+$-T cells, the average distance is ~50 kb [48]. Multiple enhancers can control the activity of one or a group of genes [47]. Enhancer-dependent transcription requires functional contacts between enhancers and target promoters. The dominant ‘‘looping’’ model suggests that active enhancers form direct physical contacts with promoters, creating a loop of intervening DNA [49–51]. This looping model has been confirmed for various enhancers [51, 52]. Binding of transcription factors to enhancers leads to the activation of transcription through recruitment of co-activators, releasing paused RNA polymerase 2 (RNAPII) and stimulating elongation [53].

Enhancers contain CpG sites and are sensitive to DNA methylation [31, 32, 36, 38, 54]. In addition to the traditionally defined enhancers that typically do not reside in CGI [4], new research has suggested that an additional class of strong enhancers exists within the >10% of CGI that are not located near known genes [44]. Further, in research focused on understanding how DNA methylation associates with gene expression (discussed below in "DNA methylation and gene expression"), enhancers are frequently enriched among the CpG sites at which methylation significantly correlates with gene expression. This has led to a proposed model of gene regulation wherein promoter methylation is relatively static, being either constitutively hypermethylated (silenced) or hypomethylated (permissive), and dynamic enhancer methylation modulates tissue-specific gene expression levels [54].

Although population studies of disease have revealed many loci that associate with various diseases in genome-wide association studies, a minority of them fall within coding regions of the genome. Many of them, however, do fall into annotated enhancer regions. This has led to the discovery of a wide range of diseases that are associated with enhancer mutations, or interruptions in enhancer-promoter interactions (reviewed in [53, 55–58]). Aberrant DNA methylation at enhancers is also seen in various cancers [59–61]. These two points highlight the potential importance of enhancer CpG methylation in gene expression and disease.

Insulators. Although past research interpreted the role of insulators as restricting enhancer-promoter interactions, more recent research suggests that the major role of insulators in the genome is to facilitate the interaction between regulatory sequences [50, 62]. As part of the "looping" model mentioned above, insulators are thought to promote gene expression by bringing enhancers and promoters into close proximity through the binding of the CCCTC-binding factor (CTCF). CTCF binds insulator elements, can dimerize to form stable chromatin loops, and associates with the ring protein, cohesin [50, 51]. These characteristics allow two insulators to come together, forming a loop of intervening DNA stabilized by cohesin, to bring a promoter and enhancer into relatively close proximity [49, 51]. In a high resolution study of Hi–C, a method to detect chromatin interactions in the 3D space of the mammalian nucleus, most chromatin loops were anchored by insulators, supporting the role of insulators as master regulators of chromatin organization [51]. Disruption of CTCF-mediated chromatin interactions can influence gene expression and are implicated in complex diseases, like cancer [61, 63–65]. Lastly, the binding affinity of CTCF to insulator sequences is influenced by DNA methylation, which suggests that DNA methylation at insulators may facilitate functional, biological changes in a cell [64, 66].

Non-coding RNAs. MicroRNAs (or miRNAs) are small non-coding RNA molecules about 22 bases in length. In humans, miRNA genes are translated and form large hairpin or stem-loop structures called a primary- or pri-miRNA. The pri-miRNA is cleaved to become a pre-miRNA, which is exported from the nucleus before being cleaved again by the enzyme Dicer and incorporated into the Argonaute containing RNA-induced silencing complex (RISC). The miRISC (miRNA + RISC) complex binds by sequence complementarity of the miRNA to the 3' UTR of a target gene. The miRISC complex the regulates mRNAs through one of three mechanisms: 1) through direct cleavage of the mRNA by Argonaute, 2) recruiting deadenylation factors that remove the target mRNA's poly(A)-tail, triggering degradation or 3) translational repression by interfering with translational machinery or direct proteolysis. MicroRNAs are thought to post-transcriptionally regulate more than 50% of human mRNAs [67]. This gene regulation pathway is involved in the development of complex disease [67–70].

Evidence suggests that long intergenic non-coding RNAs (lincRNAs) play an important role in gene expression, particularly as a scaffold molecule that attracts either transcriptional or histone modifying machinery [71]. In the case of the former, long non-coding RNAs (lncRNAs), called eRNAs (enhancer RNAs), are transcribed from enhancer sequences and may act as scaffolding for DNA looping or co-activator recruitment to a gene promoter [49]. Classic examples for the former case come from X-chromosome inactivation and the imprinted locus, *KCNQ1*. In X-chromosome inactivation, the lncRNA REPA recruits and directly interacts with the polycomb repressive group protein catalytic subunit, EZH2 in the earliest stages of X-inactivation. At the *KCNQ1* locus, the lncRNA *KCNQ1OT1*, directly recruits polycomb repressive group protein, PRC2, both in *cis* and in *trans*, to repress six paternally imprinted genes at the locus through H3K27me3 modification [71].

Transcription of both miRNAs and lncRNAs is sensitive to DNA methylation [34, 37, 40, 68, 69]. Their methylation-sensitive activity suggests that non-coding RNAs represent another layer of control in which DNA methylation may function to modulate gene activity.

Gene bodies. Although a great deal of research focuses on CpG sites located near gene promoters, they also exist within the bodies of genes. Though most gene bodies are

CpG-poor and methylated, they can include CpG islands and it is conventionally believed that methylation at intragenic CpGs positively correlates with gene expression [4, 54, 72]. Research outlined by Lister et al. (2009) and Zilberman et al. (2007) are commonly cited as proof of this phenomenon. Lister et al. assessed 5mC in various contexts in both a pluripotent stem cell line (H1) and in fully differentiated fetal lung fibroblasts (IMR90). They found that when looking across 5,644 different genes, gene body CpG methylation positively correlated with gene expression level. That is to say – highly expressed genes, on average, had a higher proportion of methylated intragenic CpGs [1, 73]. However, it is important to note that this method of analysis can only tell us the general trends across all genes; it does not tell us how changes to intragenic CpG methylation affect gene expression or how individual genes are epigenetically regulated.

Zilberman et al. studied 5mC in the flowering plant model system, *Arabidopsis thaliana.* They also found a correlation between levels of intragenic CpG methylation and gene expression across all studied genes. A notable but often overlooked finding in this research is that changes to gene body methylation are most often negatively correlated with gene expression. Zilberman et al. conclude that gene body methylation impedes transcriptional elongations in *A. thaliana* [73]. Their conclusion supports earlier work in mammalian cells by Lorincz et al. (2004). Newer work assessing site-specific CpG methylation and gene expression also confirms the abundance of negative methylation-expression correlations [30, 33, 35, 41, 74].

Research by Yang et al. (2014) finds both negative and positive correlations of CpG methylation and gene expression in human cells. However, they conclude that positive correlations of CpG methylation and gene expression are due to the role of CpG methylation in promoting gene expression and that any negative correlations are the result of intragenic regulatory elements, like cryptic or alternative promoters and enhancers [72]. The idea that intragenic DNA methylation can serve to regulate tissue-specific gene expression via alternative promoter methylation is fairly well supported. The primary example comes from Maunakea, et al. (2010). Using a robust combination of technologies, and confirming their results in both mouse and human, they show that intragenic CpG islands overlap alternative promoters and that this pattern is conserved across species [75]. To support their conclusion that intragenic methylation drives expression, Yang et al. provide evidence that the loss of expression that accompanies gene-body demethylation on treatment with 5-aza-2′-deoxycytidine is regained along with intragenic DNA methylation upon recovery. However, it is difficult to determine from these results if intragenic methylation is the driver or passenger of transcriptional control [72].

Jjingo et al. (2012) note that the relationship of intragenic DNA methylation and gene expression is not monotonic, but rather bell-shaped. They show that intragenic methylation increases with average transcription level, but decreases with the highest levels of transcriptional activity. This observation leads them to hypothesize that intragenic methylation can block transcription from alternate or cryptic promoters, but is largely is a passenger event of transcriptional activity. In their model, DNA methyltransferases are precluded from chromatin by the regular placement of nucleosomes in quiescent genes. The activity of RNA polymerase II (POL2) attracts DNMTs and disrupts nucleosome placement enough for methylation to occur in gene bodies. However, at very high levels of transcription, the processivity of POL2 itself precludes DNMT binding [76]. This hypothesis is supported by molecular research performed by Baubec et al. (2015), showing that *de novo* methylation is set in the gene bodies of actively transcribed genes by DNMT3B and is precluded by regularly placed nucleosomes. They go on to show that SETD2-mediated H3K36me3, which is common in the gene bodies of transcribed genes, guides *de novo* methylation by DNMT3B [77].

In all, gene body methylation is still a provocative topic. One might conclude from the evidence presented here that changes in intragenic DNA methylation primarily inhibit gene expression at alternate and cryptic regulatory elements. Further, it would appear that transcription may facilitate intragenic DNA methylation, which may lead to the false conclusion that DNA methylation fosters gene expression.

DNA-methylation-based association studies

In the above sections, I have described the pathways through which 5mC patterns are established and maintained, described the mechanisms by which 5mC influences gene expression and explored mechanisms under active research. I have presented evidence that the mark is dynamic throughout life, and also that DNA methylation patterns can and do associate with disease. In this section, I will explain how such associations are identified and interpreted, and some of the ways in which interpretation is currently limited.

In 2007, the advent of high-throughput single nucleotide polymorphism (SNP) microarrays and a greater understanding of population-based linkage disequilibrium facilitated the first genome-wide association studies (GWAS) [78]. In these studies, variation in the phenotype under investigation was compared to variation at loci across the genome in a study cohort, in order to draw statistical associations between specific genomic variants and a wide range of biological processes and diseases. In the years since, thousands of common genetic variants and single-nucleotide polymorphisms (SNPs) have been found to have strong and replicable associations with a wide range of common diseases (http://www.genome.gov/gwastudies/) [78].

Assessment of genome-wide DNA methylation. Quantifying DNA methylation levels is not as straightforward as assessing genomic content, since the underlying sequence is, by definition, the same. A second limitation is that distinct tissue types have differing DNA methylation profiles, which can lead to spurious associations due to statistical confounding. Therefore, studies of DNA methylation, unlike genetic studies, must be performed in an isolated tissue type. There are three primary approaches to assessing DNA methylation, each with its own strengths and weaknesses. Earlier methods relied on pairs of restriction enzymes where one is methylation sensitive and the other is not, but both recognize the same sequence [79]. When paired with polymerase chain reaction (PCR), this method is useful for assessing one or a few CpG sites or assessing global changes in DNA methylation. The use of restriction enzymes has since been extended to use with microarrays and DNA sequencing, but the resolution of this method is restricted to only those CpG sites in the correct context for appropriate restriction enzymes [78, 80]. The second approach, often called Methylated DNA ImmunoPrecipitation (MeDIP), uses antibodies to methylated DNA or DNA bound to MBD family proteins in an immunoprecipitation to enrich samples for methylated DNA [81, 82]. This method has also been combined with microarrays and sequencing, and is most useful in regions of the genome that tend to be intermediately methylated [78]. The third approach had its beginnings in the early 1990s when Frommer and others found that sodium bisulfite treatment converted cytosine bases, but not 5-methylcytosine bases, to uracil [83]. Bisulfite conversion has been one of the more popular approaches to studying the DNA methylome as it induces changes in sequence that then can be measured via conventional genotyping technologies. For example, it can be combined with PCR to study specific regions, or with microarrays or whole-genome sequencing to study site-specific patterns genome-wide [78, 80]. In recent years, the most widely-used application of bisulfite conversion employs a microarray, or beadchip, from Illumina to assay approximately 27k [84], 450k [85] or 850k CpG sites in the human genome quickly and in parallel. The introduction of this technology facilitated high-throughput analysis for larger, population-based studies of DNA methylation.

Epigenome-wide association studies. With the advent of high-throughput methylation analyses, agnostic screening of single-CpG-resolution methylation has become possible [84]. Epigenome-wide association studies (EWAS; analogous to GWAS) are a common tool for studying the role of DNA methylation in disease. The goal of EWAS is to identify CpG sites across the genome at which changes in methylation status associate with a trait of interest.

Generally, tissue samples are collected from a study cohort or from cell culture. DNA is extracted from each sample, bisulfite converted, PCR amplified and then fragmented. Then, for every sample, at each assayed CpG site, the converted DNA is annealed to a pair of probes that complements and fluorescently reports either a CpG or a TpG. In this way, the sample-wide proportion of methylated to unmethylated CpG is assessed for each site. Methylation proportions at each CpG can then be statistically compared to the trait of interest, across samples [85, 86]. CpG sites showing statistically significant associations with the trait after adjustment for the number of comparisons are typically followed up via downstream analyses including replication, validation, network analyses and enrichment tests. Methylation changes have been implicated by EWAS in the development of complex, delayed-onset diseases including diabetes [87], inflammatory bowel disease [88–90], rheumatoid arthritis [91], systemic lupus erythematosus [92, 93] and many cancers [9, 93–95]. Unfortunately, despite a growing number of EWAS, we are still far from understanding how epigenetic changes contribute to the onset of complex diseases [4, 78].

Limitations of EWAS. EWAS often return large sets of marginally significant or near-significant results, many of which lie outside of defined genomic regions (i.e. genes) [88, 89]. Inferring a functional consequence of such results is difficult because our understanding of the role of methylation in gene expression and disease is incomplete. Canonically, methylation in CpG island promoters inhibits the initiation of gene transcription [4]. Interpreting the effect of EWAS hits outside promoters is difficult, as the role of DNA methylation in these regions is not yet defined [4]. EWAS results are often interpreted based on proximity to gene or presence in gene promoters alone. Some research has already suggested that enhancer-promoter relationships are not solely based on proximity and that there can be more than one promoter between an associated enhancer-promoter pair [96]. Proximity-only interpretation disregards the role of distal CpG methylation in regulating gene expression, altogether, as well as other functional relationships [86, 97, 98]. The development of methods to facilitate interpretation of EWAS is warranted and new methods are on the horizon.

DNA methylation and gene expression

Underscoring the importance of the previous section, recent research indicates that gene expression can be affected by DNA methylation at CpG sites distal (>50 kb or on a different chromosome) to the gene promoter [30, 33, 35, 41, 54]. Some of these studies tested only for expression-associated CpGs (eCpGs) within a set distance from each gene [35, 41, 54], while others sought to identify genome-wide eCpGs for each gene [30, 33]. These studies found that eCpGs are enriched in some cancer EWAS results and are likely associated with distal regulatory sequences (specifically, enhancers), which suggests that DNA methylation distal to gene promoters may play a role in disease by altering gene expression [54, 99].

Unfortunately, these existing studies have been underpowered [30, 33, 41, 54], aimed at a sparse subset of the methylome [30, 33], or focused on only proximal CpGs [35, 41, 54]. Further, despite promising results indicating that eCpGs are enriched in enhancers, none of these studies have sought to address the role of other methylation-sensitive regulatory machinery in gene expression (i.e., insulators and non-coding RNAs). Lastly, while many of these studies report enrichment of eCpGs within gene bodies, none have adequately addressed the problems discussed above regarding experimental design (methylation associations across all genes vs associations for one gene across samples), the role of intragenic regulatory sequences, or the presence of overlapping genes.

Genome-wide studies of DNA methylation provide insight into its environmental response and regulatory potential

The following chapters introduce two studies: 1) an EWAS that assesses the effects of radiation on the epigenome and 2) a study that explores the myriad of ways that DNA methylation affects gene expression. The first study begins with a detailed review of the current state of research in acute and chronic epigenomic effects of ionizing radiation (Chapter II). It is followed by a research study that addresses gaps in understanding listed in the previous chapter and sheds new light on the subject (Chapter III). This research is important for its implications for space-travel and radio-therapy, as well as its ability to highlight the role of environment in epigenetic patterns. The second research study (Chapter IV) highlights the shortcomings of current EWAS interpretation, explores the roles of DNA methylation in various regulatory features in modulating gene expression, and provides a dataset to be utilized in the interpretation of existing and future EWAS in human blood cells. Finally, in Chapter V, I discuss common insights drawn from my two seemingly different studies and how these insights can be used to drive the continued success of epigenomics.

References

1. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
2. Li E, Zhang Y. DNA Methylation in Mammals. Cold Spring Harb Perspect Biol. 2014;6:a019133.
3. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25:1010–22.
4. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.
5. Bjornsson HT. Intra-individual Change Over Time in DNA Methylation With Familial Clustering. JAMA J Am Med Assoc. 2008;299:2877.
6. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 2014;15:R73.
7. Huidobro C, Fernandez AF, Fraga MF. Aging epigenetics: causes and consequences. Mol Aspects Med. 2013;34:765–81.
8. Fraga MF. Genetic and epigenetic regulation of aging. Curr Opin Immunol. 2009;21:446–53.
9. Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. Trends Genet TIG. 2007;23:413–8.
10. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22:R7–15.

11. O'Hagan HM. Chromatin modifications during repair of environmental exposure-induced DNA damage: a potential mechanism for stable epigenetic alterations. Environ Mol Mutagen. 2014;55:278–91.
12. Hu W, Pei H, Li H, Ding N, He J, Wang J, et al. Effects of shielding on the induction of 53BP1 foci and micronuclei after Fe ion exposures. J Radiat Res (Tokyo). 2014;55:10–6.
13. Spector S, Higham JES, Doering A. Beyond the biosphere: tourism, outer space, and sustainability. Tour Recreat Res. 2017;42:273–83.
14. Azzam EI, Toledo SM de, Harris AL, Ivanov V, Zhou H, Amundson SA, et al. The Ionizing Radiation-Induced Bystander Effect: Evidence, Mechanism, and Significance. In: Sonis ST, Keefe DM, editors. Pathobiology of Cancer Regimen-Related Toxicities. Springer New York; 2013. p. 35–61. http://link.springer.com.proxy.library.emory.edu/chapter/10.1007/978-1-4614-5438-0_3. Accessed 20 May 2013.
15. Durante M, Cucinotta FA. Heavy ion carcinogenesis and human space exploration. Nat Rev Cancer. 2008;8:465–72.
16. Schimmerling W. The Space Radiation Environment: An Introduction. Health Risks Extraterr Environ. 2011. https://three.jsc.nasa.gov/concepts/SpaceRadiationEnviron.pdf. Accessed 21 Sep 2017.
17. Plante I, Ponomarev AL, Cucinotta FA. Calculation of the energy deposition in nanovolumes by protons and HZE particles: geometric patterns of initial distributions of DNA repair foci. Phys Med Biol. 2013;58:6393–405.

18. Cucinotta FA, Durante M. Cancer risk from exposure to galactic cosmic rays: implications for space exploration by human beings. Lancet Oncol. 2006;7:431–5.
19. Lebel EA, Rusek A, Sivertz MB, Yip K, Thompson KH, Tafrov ST. Analyses of the Secondary Particle Radiation and the DNA Damage It Causes to Human Keratinocytes. J Radiat Res (Tokyo). 2011;52:685–93.
20. Klaunig JE, Wang Z, Pu X, Zhou S. Oxidative stress and oxidative damage in chemical carcinogenesis. Toxicol Appl Pharmacol. 2011;254:86–99.
21. O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, et al. Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. Cancer Cell. 2011;20:606–19.
22. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. Nat Rev Cancer. 2011;11:726–34.
23. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14:204–20.
24. Bird AP, Wolffe AP. Methylation-induced repression--belts, braces, and chromatin. Cell. 1999;99:451–4.
25. Sarraf SA, Stancheva I. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol Cell. 2004;15:595–605.
26. Feng Q, Zhang Y. The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. Genes Dev. 2001;15:827–32.

27. Zhang Y, LeRoy G, Seelig HP, Lane WS, Reinberg D. The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. Cell. 1998;95:279–89.

28. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. Cell. 2007;130:77–88.

29. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010;464:1082–6.

30. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12:R10.

31. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144:327–39.

32. de Andrés MC, Imagawa K, Hashimoto K, Gonzalez A, Roach HI, Goldring MB, et al. Loss of methylation in CpG sites in the NF-κB enhancer elements of inducible nitric oxide synthase is responsible for gene induction in human articular chondrocytes. Arthritis Rheum. 2013;65:732–742.

33. Eijk KR van, Jong S de, Boks MP, Langeveld T, Colas F, Veldink JH, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics. 2012;13:636.

34. Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol. 2013;14:699–712.

35. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. Am J Hum Genet. 2013;93:876–90.

36. Kaaij LT, van de Wetering M, Fang F, Decato B, Molaro A, van de Werken HJ, et al. DNA methylation dynamics during intestinal stem cell differentiation reveals enhancers driving gene expression in the villus. Genome Biol. 2013;14:R50.

37. Kaikkonen MU, Lam MTY, Glass CK. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res. 2011;90:430–40.

38. Kozlenkov A, Roussos P, Timashpolsky A, Barbu M, Rudchenko S, Bibikova M, et al. Differences in DNA methylation between human neuronal and glial cells are concentrated in enhancers and non-CpG sites. Nucleic Acids Res. 2014;42:109–27.

39. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14:288–95.

40. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.

41. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15:R37.

42. Kumar V, Westra H-J, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, et al. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. PLoS Genet. 2013;9:e1003201.

43. Popadin K, Gutierrez-Arcelus M, Dermitzakis ET, Antonarakis SE. Genetic and epigenetic regulation of human lincRNA gene expression. Am J Hum Genet. 2013;93:1015–26.

44. Bell JSK, Vertino PM. Orphan CpG islands define a novel class of highly active enhancers. Epigenetics. 2017;12:449–64.

45. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.

46. Ritter DI, Dong Z, Guo S, Chuang JH. Transcriptional Enhancers in Protein-Coding Exons of Vertebrate Developmental Genes. PLoS ONE. 2012;7. doi:10.1371/journal.pone.0035202.

47. Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol. 2014;21:210–9.

48. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012;22:490–503.

49. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform. 2016;17:980–95.

50. Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12:283–93.

51. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

52. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell. 2014;14:762–75.

53. Erokhin M, Vassetzky Y, Georgiev P, Chetverina D. Eukaryotic enhancers: common features, regulation, and participation in diseases. Cell Mol Life Sci. 2015;72:2361–75.

54. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14:1–14.

55. Corradin O, Cohen AJ, Luppino JM, Bayles IM, Schumacher FR, Scacheri PC. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. Nat Genet. 2016;48:1313.

56. Luo Z, Lin C. Enhancer, epigenetics, and human disease. Curr Opin Genet Dev. 2016;36:27–33.

57. Matharu N, Ahituv N. Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. PLOS Genet. 2015;11:e1005640.

58. Yao L, Berman BP, Farnham PJ. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. Crit Rev Biochem Mol Biol. 2015;50:550–73.

59. Bell RE, Golan T, Sheinboim D, Malcov H, Amar D, Salamon A, et al. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. Genome Res. 2016;26:601–11.

60. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. 2016;17. doi:10.1186/s13059-016-0879-2.

61. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 2014;24:1421–32.

62. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.

63. Dong X, Li C, Chen Y, Ding G, Li Y. Human transcriptional interactome of chromatin contribute to gene co-expression. BMC Genomics. 2010;11:704.

64. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016;529:110–4.

65. Kang JY, Song SH, Yun J, Jeon MS, Kim HP, Han SW, et al. Disruption of CTCF/cohesin-mediated high-order chromatin structures by DNA methylation downregulates PTGS2 expression. Oncogene. 2015. doi:10.1038/onc.2015.17.

66. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol Cell. 2017;66:711–720.e3.

67. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet. 2012;13:271–82.

68. Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. Nat Rev Cancer. 2015;15:321–33.

69. Long X-R, He Y, Huang C, Li J. MicroRNA-148a is silenced by hypermethylation and interacts with DNA methyltransferase 1 in hepatocellular carcinogenesis. Int J Oncol. 2014;44:1915–22.

70. Sayed D, Abdellatif M. MicroRNAs in Development and Disease. Physiol Rev. 2011;91:827–87.

71. Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet. 2015;16:71–84.

72. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. Cancer Cell. 2014;26:577–90.

73. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 2007;39:61–9.

74. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, et al. Methylomics of gene expression in human monocytes. Hum Mol Genet. 2013;22:5065–74.

75. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466:253–7.

76. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. Oncotarget. 2012;3:462–74.

77. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015;520:243–7.

78. Bell CG. Epigenome-Wide Association Studies: Potential Insights into Human Disease. In: Naumova AK, Greenwood CMT, editors. Epigenetics and Complex Traits. Springer New York; 2013. p. 287–317. http://link.springer.com.proxy.library.emory.edu/chapter/10.1007/978-1-4614-8078-5_13. Accessed 28 Apr 2014.

79. Bird AP, Southern EM. Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis. J Mol Biol. 1978;118:27–47.

80. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet. 2008;9:179–91.

81. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005;37:853–62.

82. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell. 2006;126:1189–201.

83. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89:1827–31.

84. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1:177–200.

85. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288–95.

86. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013;10:949–55.

87. Toperoff G, Aran D, Kark JD, Rosenberg M, Dubnikov T, Nissan B, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. Hum Mol Genet. 2012;21:371–83.

88. Harris RA, Nagy-Szakal D, Pedersen N, Opekun A, Bronsky J, Munkholm P, et al. Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. Inflamm Bowel Dis. 2012;18:2334–41.

89. Lin Z, Hegarty J, Cappel J, Yu W, Chen X, Faber P, et al. Identification of disease-associated DNA methylation in intestinal tissues from patients with inflammatory bowel disease. Clin Genet. 2011;80:59–67.

90. Lin Z, Hegarty JP, Yu W, Cappel JA, Chen X, Faber PW, et al. Identification of disease-associated DNA methylation in B cells from Crohn’s disease and ulcerative colitis patients. Dig Dis Sci. 2012;57:3145–53.

91. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.

92. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res. 2010;20:170–9.

93. Richardson BC. Role of DNA methylation in the regulation of cell function: autoimmunity, aging and cancer. J Nutr. 2002;132 8 Suppl:2401S–2405S.

94. Ahuja N, Issa JP. Aging, methylation and cancer. Histol Histopathol. 2000;15:835–42.

95. Liu L, Wylie RC, Andrews LG, Tollefsbol TO. Aging, cancer and nutrition: the DNA methylation connection. Mech Ageing Dev. 2003;124:989–98.

96. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016;48:488–96.

97. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13:705–19.

98. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.

99. Pandiyan K, You JS, Yang X, Dai C, Zhou XJ, Baylin SB, et al. Functional DNA demethylation is accompanied by chromatin accessibility. Nucleic Acids Res. 2013;41:3973–85.

Chapter II. Epigenetic Memory of Space Radiation Exposure

Elizabeth M. Kennedy, Karen N. Conneely and Paula M. Vertino

Adapted from an original publication, as cited:

Kennedy EM, Conneely KN, Vertino PM. Epigenetic Memory of Space Radiation Exposure. The Health Risks of Extraterrestrial Environments. 2014. https://three-jsc.nasa.gov/articles/Vertino.pdf.

Manned interplanetary travel is imminent, but currently inhibited not only by factors such as technology and budget, but also by the uncertainty surrounding the health risks associated with galactic cosmic radiation (GCR) exposure, for which there is currently no effective means of shielding. Much research has focused on the influence of high energy and charge (HZE) nuclei, the most detrimental component of GCR, on the genome. Exposure to radiation, either terrestrially or in space, induces DNA damage that, if not accurately repaired, can give rise to genetic mutations with long-term biological implications including development of chronic diseases, such as cardiovascular disease and cancer. The radiation encountered on Earth is composed primarily of low linear energy transfer (low-LET) photons (eg. gamma rays or X-rays) that react only sparsely with cellular components, with DNA damage arising primarily from free radicals generated by the ionization of nearby water molecules. In contrast, HZE ions have a high linear energy transfer (high-LET) and deposit energy linearly along the particle trajectory, interacting directly with macromolecules in addition to giving off high energy electrons (δ rays) that extend laterally for several microns. This creates a more dense ionization track as the particle moves through the cell and nucleus and results in a more complex DNA damage that involves mixtures of more than one type of DNA damage in close proximity (double strand breaks, single strand breaks, base damage, etc) [1, 2]. This clustering of multiple types of DNA damage is thought to pose a challenge for the DNA repair machinery, potentially accounting for the more deleterious biological effects of high-LET radiation than low-LET radiation at similar doses. High-LET radiation exposure can also have a lasting impact on cellular physiology without any DNA mutation, via alterations in DNA methylation, though the frequency and consequence of these changes has not been well-documented.

DNA methylation

DNA methylation refers to a chemical modification of cytosine bases in DNA. In most cells this modification occurs at adjacent cytosine-guanine nucleotides called CpG sites. DNA methylation influences which genes are expressed and in what context, allowing for a diverse collection of cell types with distinct functions to arise from a single common genome. DNA methylation is thus considered an “epigenetic” modification, as it represents another level of information that sits “on top of” the genetic code (the DNA sequence) and influences how the genetic code is interpreted [3]. Most CpG sites in the genome are methylated, including the regions within and between genes. Methylation in these intergenic regions plays an important role in the silencing of mobile elements (transposons) and in maintaining chromosome structure and stability. Also embedded within these methylated regions are ‘enhancers’, long-range regulatory elements that can influence the expression of genes from a great distance. In contrast, a small fraction of the genome is composed of areas dense with CpG sites, called CpG ‘islands’, that typically remain unmethylated in normal cells. More than half of human protein coding genes and many non-coding RNAs are regulated by promoters that lie within CpG islands. A small subset of CpG island promoters acquire methylation normally during development and cell-type specification as part of a program to ensure the long-term silencing of the neighboring gene (for example, during the programmed silencing of one X-chromosome in female mammals). Such epigenetic regulation of gene expression is critical to normal development and plays an important role in the maintenance of cellular identity. However, unlike DNA sequence, DNA methylation patterns change readily over time and with normal aging, and represent an important feature of how organisms adapt to a changing environment [4].

DNA methylation patterns can change in response to extrinsic environmental factors (e.g. nutrition, chemical pollutants) and with age [5–7]. Because DNA methylation patterns are copied along with the DNA sequence during cellular replication, induced changes to the epigenetic state will persist over multiple cell divisions, resulting in a lasting and mitotically heritable "memory" of prior exposures. Such induced alterations to DNA methylation have the potential to contribute to the long-term health risks associated with radiation exposure. The consequence of DNA methylation changes will depend on their genomic context. For example, loss of methylation outside of CpG islands occurs during aging and cancer progression and can lead to aberrant activation of transposons, chromosome instability, and (potentially) to the activation of cryptic enhancers [8]. Methylation in the transcribed regions of genes is positively correlated with gene expression, and depletion in these areas could result in the reduced expression of some genes [9]. Conversely, the aberrant gain of methylation in normally unmethylated CpG islands is associated with gene silencing and has been shown to contribute to cancer formation through the stable and heritable silencing of important growth suppressor genes [3].

Effects of radiation on DNA methylation

Most research to date assessing the impact of radiation exposure on the epigenome has focused on the effects of low-LET X-rays on global methylation trends. Nearly all of these studies report global hypomethylation in response to relatively high doses of X-rays (eg. up to 10 Gy) [10–15]. It has been proposed that the global hypomethylation is a consequence of the inability of the maintenance DNA methyltransferase, DNMT1, to keep up with the newly synthesized DNA generated during the repair of massive DNA damage [15], although decreased expression of DNMT1 itself and increased expression of miR-29, which negatively regulates methyltransferase expression, have also been reported [10, 12].

Considering the unique characteristics of the high-LET radiation track and the damage it elicits, there is the potential for unique effects on the epigenome. Indeed, current research indicates that exposure to high-LET radiation can also result in lasting changes in the total levels of DNA methylation in the genome, and that those changes may be different from the DNA methylation changes seen in response to equivalent doses of low-LET radiation [11, 13, 16, 17]. Although there is some disagreement among the few studies that have assessed the effects of high-LET and HZE radiation on the epigenome, the majority indicate a trend toward global hypermethylation [10, 13, 16, 17].

Outstanding questions

Does radiation exposure cause DNA methylation changes at specific regions of the human genome, like genes or other genomic compartments? Are the methylation changes that occur with radiation exposure random, or are there regions that are more prone to radiation-induced methylation 'damage'? Few studies have assessed the effect of high-LET radiation at specific CpG sites. While CpG sites in the promoters of a few select genes were tested in the above studies, no consistent alterations were observed. Methods are now in place to study methylation at essentially all 28 million CpG sites [9], allowing the complexity and target specificity of CpG methylation changes across the entire human genome to be explored.

Given that DNA damage can precipitate local changes in DNA methylation [3, 18, 19] and DNA methylation changes are known to occur with aging, and to contribute to the progression of cancer and other diseases [20], it is tempting to speculate that such epigenetic ‘scars’ could contribute to the long-term effects of radiation exposure even if the initial damage to the DNA is ultimately repaired (Figure 2-1). Support for this idea stems from studies by O’Hagan [19] and Morano [18] who showed that DNA double stranded breaks, which are characteristic of high-LET radiation, can result in stable DNA methylation-mediated transgene silencing after successful break repair. Alternatively, the observed changes in DNA methylation may reflect an indirect consequence of the broader cellular stress response. Indeed, exposure to reactive oxygen species associated with chronic inflammatory states are known to contribute to cancer risk, and can precipitate lasting changes in DNA methylation and chromatin modifications [21, 22]. Elevated reactive oxygen species can persist for up to two weeks after high LET radiation exposure [23] and could at least in principle contribute to an altered epigenome (Figure 2-1).

Future perspectives

While much has been learned about the functions and significance of DNA methylation over the last several decades, our current understanding is undergoing rapid revision. The recent discovery that endogenous methylated cytosine residues in mammalian DNA are naturally subject to enzyme-driven oxidation and detection of the oxidized methylcytosine derivatives (hydroxymethylcytosine (hmC), formylcytosine (fC), and carboxycytosine (caC)) in human and mouse DNA has prompted considerable effort to decipher the relative roles of these modified residues in gene expression and genome function [24]. The fC and caC forms serve as substrates for the DNA repair machinery, which removes the oxidized base and replaces it with unmodified cytosine, resulting in a net “demethylation” event, suggesting that DNA methylation patterns may be far more dynamic than previously anticipated. How radiation exposure of any type influences the formation or removal of these modified methylcytosine bases has not yet been explored. Furthermore, DNA methylation patterns are only one component of a broader epigenetic ‘code’ that includes posttranslational modifications to the histone proteins on which the DNA is wound into chromatin. Together, the pattern of DNA methylation and histone modifications helps to organize the genome into domains of different transcriptional potential. While local changes in histone modification status at the site of radiation-induced double strand breaks is known to play an important role in alerting the DNA repair machinery and cell cycle to the presence of DNA damage [25], recent work has suggested that radiation exposure (X-ray) can also influence global levels of various modified histones, suggesting a broader reprogramming of the epigenomic landscape [26]. Methods exist to map the genome-wide patterns of histone modifications as well, and have not yet been applied to high LET radiation exposure. Indeed, it has been argued that an assessment of DNA methylation and other epigenetic modifications should be considered in environmental toxicity and risk assessment, and that an understanding of the specific epigenetic “footprint” left by various chemical or physical toxins could one day be used to monitor a person’s exposure history [7]. The identification of specific and unique DNA methylation changes associated with high-LET radiation and known to be associated with diseases such as cancer, could in principle be used by NASA for ‘biodosimetry’; monitoring the biological impact of cumulative high-

LET radiation exposure and the associated health risks encountered by astronauts in deep space [27].

The technology and informatics to address our unanswered questions are available and affordable. Array-based platforms capable of analyzing >480,000 CpG sites (Illumina Human Methylation 450K array) have been available for several years now and have been applied to the study of DNA methylation during aging and in various disease states including cancer [28–31]. DNA methylation profiles are now available for thousands of human tumors from more than 25 different tumor types as part of the NIH-funded the Cancer Genome Atlas (TCGA) Project (https://tcga-data.nci.nih.gov/tcga/) and other similar international consortia (the International Cancer Genome Consortium; https://dcc.icgc.org/) and a wealth of information exists in the public domain for comparative studies. Next-generation sequencing-based methods allow for the analysis of DNA methylation at single base pair resolution, allowing for even greater genome coverage. For all of these approaches, robust statistical methods are necessary and are continuously being developed to address potential issues such as proper normalization strategies [32], cell type or tissue heterogeneity [33], adjustment for other confounding factors such as age or population stratification [34, 35], or the low number of replicates typically available in sequencing studies [36] . The questions we seek to answer about the effects of GCR involve larger questions about the functions of methylation in our genome. As we move forward to answer these questions we will not only step toward a future of space travel, but toward a greater understanding of our own biology.

Figure

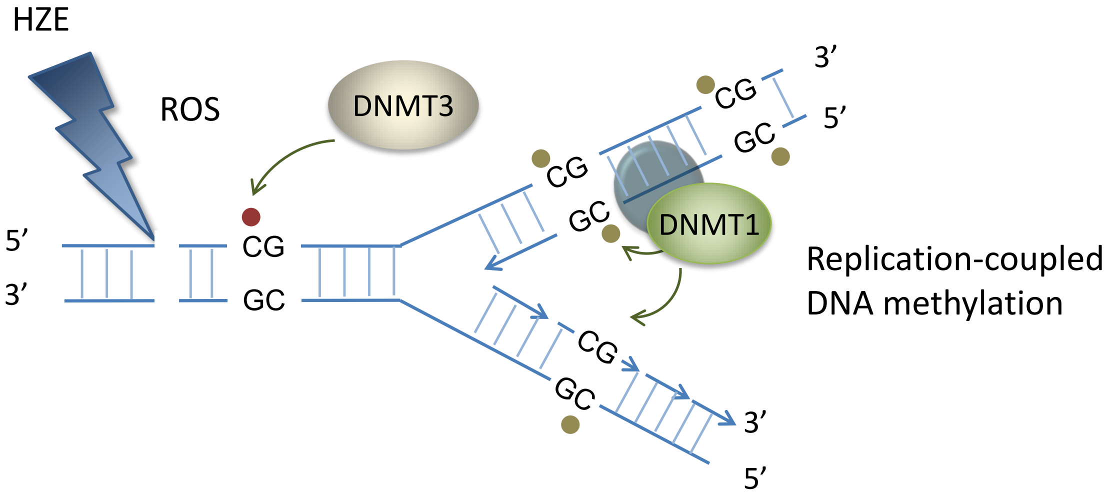


Figure 2-1 – Epigenetic alterations provide a long-term memory of prior space radiation exposure. Alterations in DNA methylation resulting from acute radiation exposure and ensuing DNA damage or persistent reactive oxygen species (ROS) have the potential to become "fixed" if they are subsequently replicated, leading to heritable epigenetic re-programming. DNA methylation (gold dots) occurs primarily at CG residues in the genome. The pattern of DNA methylation is copied during DNA replication by DNA methyltransferase 1 (DNMT1) that recognizes the methylated CG on the parental strand and transfers a methylgroup to the cytosine on the newly-synthesized strand, thereby preserving the methylation patterns in daughter cells. Gains or losses in DNA methylation induced by acute radiation exposure will likewise be copied to subsequent cell generations in the next mitosis. DNA damage, such as double strand breaks, can serve as a stimulus for a new methylation mark (red dot), and could leave an epigenetic "scar" even if the break is successfully repaired.

References

1. Cucinotta FA, Durante M. Cancer risk from exposure to galactic cosmic rays: implications for space exploration by human beings. Lancet Oncol. 2006;7:431–5.

2. Durante M, Cucinotta FA. Heavy ion carcinogenesis and human space exploration. Nat Rev Cancer. 2008;8:465–72.

3. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. Nat Rev Cancer. 2011;11:726–34.

4. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.

5. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. Am J Hum Genet. 2011;88:450–7.

6. Huidobro C, Fernandez AF, Fraga MF. Aging epigenetics: causes and consequences. Mol Aspects Med. 2013;34:765–81.

7. Mirbahai L, Chipman JK. Epigenetic memory of environmental organisms: A reflection of lifetime stressor exposures. Mutat Res Toxicol Environ Mutagen. 2014;764–765:10–7.

8. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 2014.

9. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

10. Antwih DA, Gabbara KM, Lancaster WD, Ruden DM, Zielske SP. Radiation-induced epigenetic DNA methylation modification of radiation-response pathways. Epigenetics. 2013;8:839–48.
11. Aypar U, Morgan WF, Baulch JE. Radiation-induced epigenetic alterations after low and high LET irradiations. Mutat Res Mol Mech Mutagen. 2011;707:24–33.
12. Filkowski JN, Ilnytskyy Y, Tamminga J, Koturbash I, Golubov A, Bagnyukova T, et al. Hypomethylation and genome instability in the germline of exposed parents and their progeny is associated with altered miRNA expression. Carcinogenesis. 2010;31:1110–5.
13. Goetz W, Morgan MNM, Baulch JE. The effect of radiation quality on genomic DNA methylation profiles in irradiated human cell lines. Radiat Res. 2011;175:575–87.
14. Ilnytskyy Y, Koturbash I, Kovalchuk O. Radiation-induced bystander effects in vivo are epigenetically regulated in a tissue-specific manner. Environ Mol Mutagen. 2009;50:105–113.
15. Pogribny I, Raiche J, Slovack M, Kovalchuk O. Dose-dependence, sex- and tissue-specificity, and persistence of radiation-induced genomic DNA methylation changes. Biochem Biophys Res Commun. 2004;320:1253–61.
16. Lima F, Ding D, Goetz W, Yang AJ, Baulch JE. High LET 56Fe ion irradiation induces tissue-specific changes in DNA methylation in the mouse. Environ Mol Mutagen. 2014;55:266–77.
17. Nzabarushimana E, Miousse IR, Shao L, Chang J, Allen AR, Turner J, et al. Long-term epigenetic effects of exposure to low doses of 56Fe in the mouse lung. J Radiat Res (Tokyo). 2014;55:823–8.

18. Morano A, Angrisano T, Russo G, Landi R, Pezone A, Bartollino S, et al. Targeted DNA methylation by homology-directed repair in mammalian cells. Transcription reshapes methylation on the repaired gene. Nucleic Acids Res. 2014;42:804–21.
19. O'Hagan HM, Mohammad HP, Baylin SB. Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island. PLoS Genet. 2008;4:e1000155.
20. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20:440–6.
21. Hahn MA, Hahn T, Lee D-H, Esworthy RS, Kim B-W, Riggs AD, et al. Methylation of polycomb target genes in intestinal cancer is mediated by inflammation. Cancer Res. 2008;68:10280–9.
22. O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, et al. Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. Cancer Cell. 2011;20:606–19.
23. Werner E, Kandimalla R, Wang H, Doetsch PW. A role for reactive oxygen species in the resolution of persistent genomic instability after exposure to radiation. J Radiat Res (Tokyo). 2014;55 suppl 1:i14–i14.
24. Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell. 2014;156:45–68.
25. Price BD, D'Andrea AD. Chromatin remodeling at DNA double-strand breaks. Cell. 2013;152:1344–54.

26. Maroschik B, Gürtler A, Krämer A, Rößler U, Gomolka M, Hornhardt S, et al. Radiation-induced alterations of histone post-translational modification levels in lymphoblastoid cell lines. Radiat Oncol Lond Engl. 2014;9:15.

27. Cucinotta FA. Space Radiation Risks for Astronauts on Multiple International Space Station Missions. PLoS ONE. 2014;9:e96099.

28. Fang F, Turcan S, Rimner A, Kaufman A, Giri D, Morris LGT, et al. Breast cancer methylomes establish an epigenomic foundation for metastasis. Sci Transl Med. 2011;3:75ra25.

29. Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet. 2014;23:1186–201.

30. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.

31. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010;17:510–22.

32. Wu MC, Joubert BR, Kuan P, Haberg SE, Nystad W, Peddada SD, et al. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. Epigenetics. 2014;9:318–29.

33. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

34. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinforma Oxf Engl. 2012;28:1280–1.

35. Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, et al. Accounting for population stratification in DNA methylation studies. Genet Epidemiol. 2014;38:231–41.

36. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42:e69.

Chapter III. Galactic Cosmic Radiation Induces Stable Epigenome Alterations Relevant To Human Lung Cancer

E.M. Kennedy[#], D.R. Powell[#], Z. Li, J.S.K. Bell, B.G. Barwick, H. Feng, M.R. McCrary, B. Dwivedi, J. Kowalski, W.S. Dynan, K.N. Conneely and P.M. Vertino

# These authors contributed equally

Adapted from a manuscript currently under review in *Scientific Reports*

Introduction

The potential for human interplanetary travel and deep space excursion are currently limited by concerns surrounding the long–term human health risks associated with galactic cosmic ray (GCR) exposure [1–4]. These risks include degenerative effects on the cardiovascular and central nervous systems and the risk of cancer at sites such as the lung, colon, breast, and stomach [5]. Given the absence of direct epidemiologic data, GCR exposure risk estimates currently rely on modeling based on data from ground-based experiments with cells and animals.

Terrestrial radiation is composed primarily of low linear energy transfer (low-LET) photons (e.g. γ-rays or X rays) that are sparsely ionizing and deposit energy in a dispersed manner in tissue. In contrast, the GCR spectrum is composed of hydrogen, helium and heavier atomic nuclei with high charge and energy (HZE), including ^{28}Si, ^{56}Fe, and other ions. Though a low fraction overall, these heavy particles are of particular concern as they have high linear energy transfer (high-LET) values and leave a concentrated track composed of a densely ionizing, nanometer-scale core and a penumbra of high-energy secondary electrons (δ rays) that can extend laterally for several microns as they traverse tissue [3, 6]. This creates a tightly clustered and complex mixture of DNA damage (double strand breaks, single strand breaks, base damage, etc.), which is a challenge to repair [4–7]. GCR also generates non-targeted effects in cells not directly traversed by radiation tracks (bystander effects), which may account for as much as half of the cancer risk at doses relevant to human exposure [8]. The unique biophysical properties of high-LET ions are also being exploited as a novel modality for cancer radiotherapy where the opportunity to deliver dense ionization selectively within the tumor volume has the potential to increase the efficacy for tumor control while minimizing normal tissue toxicity. Indeed, carbon-ion beam therapy is currently being evaluated for the treatment of brain tumors and other cancer types in Europe and Asia [9]. A better understanding of the biological effects of HZE particle exposure therefore has important implications for cancer causation and its treatment.

The different heavy ions that make up the GCR spectrum each have distinct effects on the gene expression patterns in cultured cells, via mechanisms that remain poorly understood [10]. These differences in gene expression may reflect modifications to the epigenome. Unlike the underlying DNA sequence, the epigenome, collectively represented in the local patterns of DNA cytosine methylation, posttranslational modifications of histones, nucleosome positioning, and long-range chromatin organization, can change readily over time and may represent an important feature of how organisms adapt to a changing environment [11, 12]. In particular, DNA methylation, which occurs primarily at cytosines in the context CG (CpG), is propagated at each cell division by the action of the DNMT1-UHRF1 complex, which copies the methylation status of CpGs on the parental DNA strand to the newly-synthesized strand, a specificity imparted by a preference for hemi-methylated CpG dinucleotides [13]. Therefore, induced changes in the DNA methylation patterns have the potential to persist over multiple cell divisions, resulting in a lasting and mitotically heritable “memory” of prior exposures. Such induced alterations to the epigenome, in addition to changes to the genome, have the potential to contribute to altered gene expression programs and the long-term consequences of radiation exposure.

To date, most of the research addressing the effects of radiation exposure on the epigenome has focused on the impact of low-LET X rays. In general, these studies have reported a trend towards global hypomethylation in response to relatively high doses of X rays (i.e. up to 10 Gy) [14–17] which may result from a decrease in DNA methyltransferase levels [17–19]. A few studies have assessed the effects of HZE particle radiation on the epigenome, and have focused on overall DNA methylation levels, repetitive element DNA methylation, or the analysis of a limited number of sites in selected gene loci. The majority indicate a global trend toward hypermethylation There are currently no studies to directly compare the effects of site-specific DNA methylation changes between low-LET terrestrial radiation (X ray) or among high-LET HZE particle radiation species on a genome-wide scale.

Here we examined the effects of ^{56}Fe and ^{28}Si, two ions found in GCR, on the methylation status of over 485,000 CpG sites across the human genome. We assessed the acute impact (48 hr. post irradiation) and long-term persistence of DNA methylation changes induced by each exposure and compared that to the effects of X rays. We find that dose-dependent changes in DNA methylation are observed early and persist over time, with each insult having unique characteristics with regards to the direction, distribution, and underlying chromatin compartment affected, suggesting that these changes arise through distinct mechanisms and may have distinct biological consequences. Further, we find that the ^{56}Fe ion-induced methylation signature uniquely reflects a cancer-specific methylation pattern observed in human primary lung cancers. Together these results speak to an epigenetic ‘memory’ of space radiation exposure.

Materials and methods

Cell Line and Culture conditions. The immortalized human bronchial epithelial cell line (HBEC3- KT) was established by introducing mouse Cdk4 and hTERT into normal human bronchial epithelial cells (HBECs) [20] and were a kind gift from Dr. J.D. Minna of the University of Texas Southwestern Medical Center. Throughout this study the HBEC3-KT immortalized line was cultured in Serum-Free Keratinocyte Medium (K-SFM) supplemented with human recombinant Epithelial Growth Factor and Bovine Pituitary Extract (Life Technologies #17005-042). Triplicate biological replicates were irradiated and maintained independently. Cells were continually grown in a 5% CO_2 environment at 37°C and passaged (1:4) twice per week for three months. Cell pellets (1e6) were collected at each passage, flash frozen, and stored at -80°C for subsequent DNA extraction.

Irradiation. High-energy HZE particle irradiations were performed in Brookhaven, NY at the NASA Space Radiation Laboratory (NSRL). The X- ray (low-LET) exposures were conducted at Emory University using an X-RAD 320 biological irradiator (Precision X-Ray, North Branford, CT).

Three biological replicate cultures containing either 1x10e6 cells (for acute time point) or 2x10e5 cells (for continuous culture) in T-25 flasks were irradiated independently with 0, 0.1, 0.3 or 1.0 Gy ^{56}Fe ions (Beam energy: 600 MeV/u; dose rate for the 0.1 Gy dose was 0.1Gy/min, for the 0.3 Gy dose, 0.3 Gy/min, and for the 1.0 Gy dose, 1Gy/min.) or with 0.0, 0.3, 1.0 Gy ^{28}Si ions (Beam energy: 300 MeV/u; dose rate for the 0.3 Gy dose was 0.28 Gy/min, and for the 1.0 Gy dose, 0.63 Gy/min).. Culture flasks were positioned orthogonally to the beam using an automated flipper provided at the NSRL. X ray irradiations were performed using identical plating conditions and exposed to doses of 0 Gy and 1.0 Gy (beam energy 320 kV; dose rate ~1 Gy/min). Immediately following irradiation, all cultures were returned to a 37°C incubator for forty-eight hours before cell pellets were collected from one set of flasks (triplicates, 2 day time point), while the remaining cultures were returned to the home laboratory at Emory University and maintained in continuous culture for an additional 3 months, with biweekly subculturing and DNA collection. For each experiment, mock-irradiated controls (triplicate cultures) were seeded and handled identically, including travel to and from the NRSL facility. All triplicate cultures were maintained independently from the start of the experiment.

DNA methylation profiling. Genomic DNA isolation was conducted at the time of sample processing for subsequent methylation analysis. Triplicate cell pellets, previously held at -80°C, were processed using the All Prep DNA/RNA kit (Qiagen #80204) according to the manufacturer's instructions. The methylation status of 485,577 CpG sites was then interrogated using the Illumina Human Methylation450K platform (Illumina, San Diego, CA). DNA (1 μg) was bisulfite modified using the EZ DNA Methylation-Direct kit (Zymo Research #D5020), fragmented, amplified and hybridized to the HumanMethylation450 BeadChip array according to the manufacturer's instructions by the Emory Integrated Genomics Core Facility. All samples from a given exposure experiment were processed in parallel and on parallel chips with replicate samples randomized with respect to chip position.

Differential DNA methylation analyses. CpGassoc [21] was used to perform quality control and differential DNA methylation analyses. The ^{56}Fe ion, ^{28}Si ion, and X ray exposed cohorts were considered separately in the analyses, and each included 3 biological replicate samples for each of 24 doses and 4 time points per exposure type (^{56}Fe ions, 4 doses x 4 time points = 48 samples; ^{28}Si ions = 3 doses x 4 time points = 27 samples; X ray, 2 doses x 4 time points = 24 samples). For each CpG site, the signals from methylated (M) and unmethylated (U) bead types were quantile normalized together (using the R-package limma; [22] and then used to calculate β-values [$\beta = M/(U + M)$], which approximate the proportion of DNA methylated at each CpG. Data points with detection p-values >0.001 were set to missing, and CpG sites with missing data for >10% of samples were excluded from the analysis. Samples with a probe detection call rate <95%, or an average signal intensity <2,000 AU or <50% of the experiment-wide sample median were also excluded. This resulted in a total of 484,434 (^{56}Fe ion); 484,765 (^{28}Si ion) and 484,384 (X ray) CpGs considered in the subsequent analysis. A linear mixed effects model was applied to identify DNA methylation changes significantly associated with dose and time-after-exposure. Intra-experiment β-values were modeled as a linear function of radiation dose, with covariates adjusting for time-after-exposure, row on chip, and a random effect for chip number. The Holm (step-down Bonferroni) method was applied to correct for multiple comparisons [23]. CpG sites with a corrected p-value<0.05 were considered nominally significant and CpGs with an uncorrected p-value< 0.001 were considered moderately significant. To identify methylation changes associated with time-after-exposure independent of dose, the same strategy was used to model β-values as a function of time-after-exposure, with covariates adjusting for radiation dose, and assay variables.

Genomic annotation and meta-analyses. CpGs were annotated to the nearest UCSC CpG Island (CGI) and RefSeq (v.75) transcription start site (TSS) using custom

R/Bioconductor scripts [24]. CpGs were categorized into those overlapping a CGI, or those within 2kb upstream (5' Shore), 2kb downstream (3'Shore), or between the downstream shore and the transcription termination site (Gene Body). CpG sites not falling into one of these classes were considered as 'other/intergenic.' The distribution of all CpGs covered by the assay versus those determined to be hypermethylated or hypomethylated were plotted relative to the nearest CGI or TSS using the density function in R/Bioconductor, where the width of the CGI was scaled to the average width of UCSC CGIs.

ENCODE [25] ChIP-seq data sets derived from A549 lung cancer cells (H3K27Ac, ENCSR000AUI; H3K4me3 ENCSR000ASH; DNaseI, ENCSR136DNA) were downloaded as mapped bam files from the ENCODE project website (https://www.encodeproject.org) . The average tag densities surrounding each set of CpG sites were calculated in 20 bp bins using the GenomicRanges R package [26], and normalized to the total number of mapped reads.

ChromHMM [27] chromatin state maps derived from normal human mammary epithelial cells (ENCFF687QKV) were used to annotate each CpG site to a chromatin compartment. For clarity, both 'Strong Enhancer' (states 4 and 5) and both 'Weak Enhancer' (states 6 and 7) were merged. States 9, 10, and 11 ('Transcriptional Transition', 'Transcriptional Elongation', and 'Weak Transcription') were merged and referred to as 'Transcribed Regions'. Odds ratios were calculated based on the number of affected sites in each compartment vs. the distribution of the CpGs on the array as a whole using Fisher's Exact test.

Analysis of lung cancer TCGA data. Illumina Infinium HumanMethylation 450K methylation data for the ^{56}Fe ion- (n=935), ^{28}Si ion- (n=300) and X ray- (n=1150) affected CpG sites was extracted for 25 matched tumor normal pairs of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) from patients identified through the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). CpG probes with a detection p-value> 0.05 across the sample set were excluded, leaving 784, 938, and 237 CpG sites from the ^{56}Fe ion-, ^{28}Si ion-, and X ray-affected sites, respectively. The methylation levels (beta values) from these sites were then used in an unsupervised, hierarchical cluster analysis based on the Manhattan distance and agglomerative complete linkage. The statistical significance of each set of exposure associated CpG sites in separating lung tumors from normal was assessed versus a randomly selected set of CpG sites of the same number using a bootstrap approach. Specifically, we randomly sampled M=1,000 times the same number of CpG sites in each exposure set from among a total of 391,954 (LUAD) or 395,241 (LUSC) CpG sites on the array that remained after exclusion of low quality probes (detection p-value across all samples >0.05). For each resampling, dendrograms were constructed using the same unsupervised clustering approach and cut based on a fixed number of k=2 clusters. An association analysis was performed based on a Chi-Square test for each resampling and p-values obtained. A Monte Carlo p-value was used to compare the ability of randomly-sampled CpG sites to separate tumor and normal samples into two clusters versus the p-value obtained from the CpG sites defined by each group (^{56}Fe, ^{28}Si, X ray).

Results

The goals of this study were to define the acute impact (48 hrs.) and long-term persistence of radiation exposure on the epigenome, and to directly compare the effects of high-LET GCR components (^{56}Fe ion, 170keV/µm; ^{28}Si ion, 70keV/µm) and low-LET X rays (2keV/µm). We hypothesized that induced changes to the DNA methylation pattern would provide a lasting imprint of the acute radiation exposure with the potential to contribute to the long-term health risks, including cancer. To test this hypothesis, triplicate cultures of immortalized human bronchial epithelial cells (HBEC-3KT) [20] were exposed to high-LET radiation (^{56}Fe ion: 600 MeV/u at 0, 0.1, 0.3, 1.0 Gy; ^{28}Si ion: 300 MeV/u at 0, 0.3, 1.0 Gy) at the Brookhaven National Laboratory NASA Space Radiation Laboratory (NSRL) or to low-LET radiation (X ray: 0, 1.0 Gy) at Emory University. Samples were collected from a fraction of the exposed population after 48 hrs. and the remaining cells were maintained in continuous culture for an additional ~35 population doublings (~2.5 months). Cells were collected for genomic DNA extraction at 48 hrs., and thereafter at ~1 week intervals. Non-irradiated control cultures underwent the same handling procedures and were maintained in parallel. Triplicate cultures were kept as independent biological replicates throughout the course of the experiment. While irradiation elicited some acute cell death (~40% at the highest doses of ^{56}Fe), the majority of cells survived to the next passage and onward.

Radiation-induced changes to the epigenome are LET dependent and ion specific. The methylation status of 485,577 CpG sites was assessed for each DNA isolate (triplicate samples for each treatment dose and time-in-culture) using the Illumina Infinium HumanMethylation 450K Platform. DNA methylation levels at each CpG site are represented as a β-value that estimate the percent of methylated alleles in the DNA sample at that position. For each radiation type, we applied a linear mixed effects model to identify CpG sites where the methylation status changed significantly with the dose. This approach allows for an independent assessment of methylation changes significantly associated with radiation dose by accounting for other covariates such as time-after-exposure. We identified 935 CpG sites where the methylation status was moderately associated with dose of ^{56}Fe ions (849 hypermethylated; 86 hypomethylated, $p<0.001$); 300 sites where the methylation status was associated with ^{28}Si ion dose (158 hypermethylated, 142 hypomethylated, $p<0.001$) and 1,150 where the methylation status was associated with X ray dose (252 hypermethylated; 898 hypomethylated, $p<0.001$).

The effects of radiation on both global and site-specific CpG methylation patterns were dependent on radiation type (^{56}Fe ions, ^{28}Si ions, X rays; Figure 3-1). ^{56}Fe ion exposure tended to affect CpG sites that are normally less methylated (mean=21.9%) and to induce their hypermethylation (Figure 3-1 A, B). ^{28}Si ion exposure primarily affected CpG sites that start with intermediate DNA methylation levels and had a roughly equivalent tendency to promote their hyper or hypomethylation (Figure 3-1 C, D). X ray exposure primarily affected more highly methylated CpG sites (median = 61.9%) and led to their hypomethylation (Figure 3-1 E, F). These radiation-type specific genome-wide trends were further reflected in an average dose-dependent trend towards hypermethylation in response to ^{56}Fe ion exposure, no average change in response to^{28}Si ion exposure, and an average trend towards hypomethylation in response to X ray exposure (Figure 3-1 G-I). These results are consistent with previous findings regarding DNA methylation content in high-LET radiation exposed cells [15, 28–30] and highlight the additional information provided by site-specific CpG methylation analyses versus bulk genomic measurements of DNA methylation.

The same trends were also evident in the analysis of the individual affected sites. A heat map representation of individual CpG sites again showed the preponderance of hypermethylation events for ^{56}Fe ion-exposed cells, nearly equivalent hyper and hypomethylation observed for ^{28}Si ion-exposed cells, and hypomethylation among X ray exposed cells (Figure 3-2). As implied by the genome-wide trends, the sites affected by each radiation source showed distinct patterns. Indeed, there was little or no overlap in the specific CpG sites affected by each radiation type (two or fewer CpG sites shared in total between any two radiation types). Taken together, these data are consistent with a graded methylation response with regards to LET (^{56}Fe, 170keV/µm; ^{28}Si, 70keV/µm; X rays, 2 keV/µm) rather than a sharp distinction between high-LET heavy ions and low-LET photon (X ray) radiation.

Radiation-induced changes to the epigenome occur early and persist over time. We next considered the fate of radiation-induced DNA methylation changes over time. To focus specifically on the fate of radiation-induced methylation changes, we selected those CpG sites where the change in methylation was moderately associated with radiation dose, but was not independently associated with time-dependent methylation 'drift' (see below). This left 844 ^{56}Fe ion-affected CpG sites (768 hyper; 76 hypo), 280 ^{28}Si ion-affected sites (153 hyper, 128 hypo) and 1120 X ray-affected sites (243 hyper, 877 hypo). We determined the change in mean β-value over time, relative to non-irradiated control cells at 48 hr (the earliest time point) (Figure 3-3). Although there was some variation in methylation with time among non-irradiated cells (note that the distribution of methylation levels at the 0 Gy dose broadens over time in each panel of Figure 3-3), the dose-dependent change in DNA methylation induced by each radiation source evident two days after radiation exposure was largely retained more than 50 days later (28-34 population doublings) (Figure 3-3). These data suggest that in general the radiation-induced methylation changes occur early and persist over time, resulting in a stable and heritable change in the epigenome.

Methylation "drift" over time. As noted by others [18], we observed considerable methylation "drift" over time in cell culture, independent of radiation exposure (Supplemental Figure S3-1A). Indeed, application of the linear mixed effects model identified thousands of sites significantly associated with time-after-exposure, independent of dose (i.e., when dose was considered as a covariate). For each exposure type, >2,900 sites were significant after Bonferroni (Holm) adjustment ($p<1e-7$) and >77,000 CpG sites were significant according to an FDR criterion ($FDR<0.05$; $p<0.01$). The average rate of change was consistent across the different experimental series performed over two years at different times, and was estimated to be 0.001% methylation per day, or the equivalent of a shift in methylation status of 1 in 1,000 DNA molecules per day. A comparison of those sites significantly associated with time from each experiment indicated that, whereas there was significant overlap in the sites affected from one series to the next, the direction of change was not always the same. Indeed, the two series that were performed in parallel and in the same time-frame (^{28}Si ion, X ray) showed the greatest concordance with respect to both the sites affected and direction of change, but were less concordant when either was compared to the ^{56}Fe ion series which occurred at a later date (Supplemental Figure S3-1B,C). Our results suggest that although many CpG sites are prone to methylation drift in cell culture, other factors appear to impact the direction of drift (i.e., hyper- vs. hypo-methylation). A comparison of those sites whose methylation state was significantly associated with both radiation dose and time (n=91 for ^{56}Fe ion; 19 for ^{28}Si ion; and 30 for X ray) showed that the effects of radiation and intrinsic drift were largely independent in that once imposed, the effect of radiation dose on individual CpGs had little impact on the rate or direction of drift over time (Supplemental Figure S3-1D). Thus, the effects of radiation appear to be superimposed upon an intrinsic tendency for methylation drift with cell division.

HZE ions of different charge and energies affect different genomic compartments. Given the largely independent subsets of CpGs affected by the different radiation types, we next sought to determine the relationship between source-specific DNA methylation changes and other genomic and epigenomic features. We examined the distribution of CpG sites significantly associated with ^{56}Fe ion, ^{28}Si ion, or X ray dose relative to genetic features, including the distribution in and around CpG islands and genes (Figure 3-4). Relative to the distribution of all probes on the array, ^{56}Fe ion-affected CpG sites, most of which were hypermethylated, tended to lie within CpG islands (which generally lack DNA methylation) and around transcription start sites (TSS). These hypermethylated sites were particularly enriched in CpG island “shore” regions (defined here as 2 kb from the 5’ or 3’ edge of the CpG island domain), while the few sites that became hypomethylated arise from outside these regions and away from CpG islands. In contrast, ^{28}Si ion-affected sites tended to be depleted in CpG islands and shores and instead were enriched among gene bodies and other distal regions. Overall, X ray-affected sites were distributed similarly to the probes on the array. The majority of sites that were hypomethylated were located in genomic regions outside of CpG islands, which are typically methylated; the few sites that were hypermethylated were enriched in CpG islands, which are typically unmethylated.

To further investigate the genomic compartments affected by radiation-induced methylation changes, we examined the relationship to chromatin features. Using genome-wide ChIP-seq data for histone modifications, RNA polymerase occupancy, and other chromatin features, Ernst, et al. (2011) used a hidden Markov model to partition the genome into functional domains, termed ChromHMM. We analyzed the ChromHMM states and existing genome-wide datasets to evaluate the chromatin structure surrounding the radiation-sensitive CpG sites. This analysis revealed that the ^{56}Fe ion-affected sites were more likely to occur in areas with a more "open" chromatin structure, including promoters and enhancers (Odds Ratio=1.3-1.5 fold; $p<0.004$), but were depleted from the transcribed regions of genes (Odds Ratio=0.47, p=2.8E-13; see Figure 3-5). Consistent with a propensity for enhancers, ^{56}Fe ion-affected sites were enriched in regions that are accessible to DNase I and marked by acetylated histone H3 lysine 27 (H3K27ac), a mark of active enhancers, relative to all sites on the array and as compared to the ^{28}Si ion- or X ray-affected sites, which were depleted in these features. In contrast, ^{28}Si ion-affected sites were depleted in genes and features of active/accessible chromatin (i.e. H3K27Ac, DNaseI accessibility, H3K4me3) and were more likely to occur in repressed chromatin environments (i.e. sites marked by heterochromatin and polycomb; Odds Ratio=1.5-1.6, $p<0.02$; Figure 3-5). X ray-affected sites were enriched in transcribed regions (Odds Ratio=1.3, $p<0.001$) (consistent with an enrichment in gene bodies shown above), but relatively depleted in features of active/accessible promoters and enhancers. Taken together, these data suggest that different sources of radiation preferentially affect sites in different chromatin contexts (e.g. enhancer, promoters, condensed chromatin), which could underlie their distinct biological consequences.

Methylation status of ^{56}Fe ion-affected CpG sites distinguishes primary lung tumor from normal tissue. The above data indicate that particles of different qualities and energies have unique impacts on the epigenome, which may ultimately manifest in distinct biological consequences. We next sought to determine the relevance of these radiation-induced CpG methylation changes to human lung cancer. We leveraged the human epigenome information available from hundreds of primary lung tumors that have been analyzed on the HumanMethylation450K platform as part of the Cancer Genome Atlas (TCGA) Project. Level 3 DNA methylation data (β-values) were extracted for the ^{56}Fe ion- (n=935), ^{28}Si ion- (n=300) and X ray- (n=1150) sensitive CpG sites for a set of 18 tumor-normal pairs of human lung adenocarcinoma (LUAC) and 7 tumor-normal pairs of squamous cell carcinoma of the lung (LUSC). The methylation status of these sites was then used in an unsupervised cluster analysis (complete linkage clustering, Manhattan distance). Interestingly, the methylation status of the ^{56}Fe ion sites, in particular, cleanly separated primary tumor specimens from normal tissue among the LUAC samples (p= 1.46e-08) as well as the LUSC samples (p=0.0013), whereas neither the ^{28}Si ion-affected CpG sites nor the X ray- affected CpG sites showed any significant association (Figure 3-6). To test the robustness of the separation achieved by the methylation at the ^{56}Fe ion-affected sites, the clustering approach was repeated 1,000 times using an equivalent number of CpG sites (n=777, LUAD; n=782, LUSC) chosen at random from a total of ~390,000 CpG sites with a detection p-value across all TCGA samples of <0.05. None achieved significance greater than the ^{56}Fe ion signature sites for either LUAD (random sampling p-value range 0.0005-0.0850; median = 0.009) or LUSC (random sampling p-value range 0.0590-0.1069; median 0.0885). Thus, the methylation status of CpG sites sensitive to ^{56}Fe ion exposure in human bronchial epithelial cells is uniquely characteristic of human lung cancer.

Discussion

Ionizing radiation (IR), such as γ- or X rays, increases the age-related risk of many common human cancers, with lung cancers representing about a third of cases linked to prior radiation exposure among atomic bomb survivors and occupational exposures in nuclear reactor workers [5, 31, 32]. In the absence of epidemiologic data on humans exposed to GCR, current estimates of cancer risk are based primarily on animal models, which have shown that exposure to high-LET radiation sources results in a greater tumorigenic potential and a more aggressive phenotype (e.g. shorter latency, accelerated progression, and increased metastatic potential) as compared to low-LET sources [33–37]. However, the degree to which these cancer risk estimates can be directly extrapolated to astronauts and space radiation exposure is fraught with uncertainties, due in part to incomplete understanding of the biological impact of high-LET radiation exposure and how it differs from terrestrial radiation sources as well as other confounding factors such as smoking status [33, 38]. Thus, biomarkers that can be used to monitor exposures and reliably predict disease risk are sorely needed.

Here we show that HZE particles induce a unique imprint on the epigenome. Significantly, we found that radiation-induced methylation changes occur early and persist over time, reflecting a stable and heritable change to the epigenome. The radiation-induced changes to DNA methylation patterns were source-dependent, and impact DNA in different chromatin contexts, implying that they arise through distinct mechanisms and may have distinct biological consequences. Although limited by the representation of CpG sites on the Illumina array, which is biased towards genic regions and excludes repetitive DNA and other regions of constitutive heterochromatin, each radiation source affected the epigenome in distinct ways. For example, ^{56}Fe ions have a propensity to affect regions of accessible chromatin such as gene promoters and distal regulatory elements ('enhancers') whereas ^{28}Si ions preferentially affected DNA in more repressed, heterochromatic regions. Whether these differences in DNA methylation are a reflection of a difference in the susceptibilities of different chromatin regions to radiation-induced DNA damage or to its repair is unknown, but DNA damage (double strand breaks (DSB), oxidative damage) has been suggested to promote the recruitment of DNA methyltransferases and other histone modifiers (e.g. PRC1/2; SIRT1) to mediate local chromatin repression that persist in subsequent cell divisions [39–43]. Interestingly, an electron microscopic study of high-LET (carbon-ion) induced DNA damage showed that unlike X-irradiation, which induced DSBs distributed throughout the nucleus that were efficiently cleared, high-LET radiation induced clustered lesions along the particle trajectory that localized primarily to electron dense heterochromatic regions [44]. These phosphor-Ku70-bound clusters grew larger over time, suggesting inefficient repair. While there is no way to directly relate our nucleotide level analysis with these broader scale observations, it is intriguing to speculate that a persistence radiation-induced lesions in heterochromatin might underlie the differences in tumor-promoting activities between high-LET and low-LET radiation [45, 46] or even that between different HZE ions [34, 36].

To probe the significance of our findings with respect to human lung cancer, we leveraged the human epigenome information available from hundreds of primary lung tumors that have been analyzed as part of the Cancer Genome Atlas (TCGA) project. We found that the methylation status of high-LET radiation sensitive CpG sites, particularly those impacted by ^{56}Fe ion exposure, could discriminate tumor from normal tissue for both lung adenocarcinomas and squamous cell lung carcinomas. No such relationship existed for the sites affected by ^{28}Si ion or low-LET radiation exposure. Thus, our results suggest that HZE particle exposure creates a DNA methylation ‘signature’ that uniquely reflects cancer-specific methylation patterns observed in human primary lung cancers. That it is the ^{56}Fe ion-affected signature in particular that is capable of segregating tumor from normal is perhaps not surprising given that the ^{56}Fe ion-affected sites are enriched in the regions surrounding CpG islands (the ‘shores’) and in accessible regions with chromatin features indicative of weak/poised promoters and enhancers; regions of the genome that exhibit the most variable levels of methylation across tissues/cell types and between individuals [47–49]. In contrast, CpGs within the CpG dense regions that encompass most promoters (CpG ‘islands’) typically remain unmethylated, and with few exceptions, maintain an open and permissive chromatin state (marked by H3K4me3; DNaseI hypersensitive) across tissues and cell types allowing for a wide-range of potential gene expression levels. Indeed, methylation of such regions is a relatively poor correlate of gene expression [50, 51]. In contrast, the regions immediately adjacent to CpG islands (the ‘shores’) and enhancer elements exhibit the greatest variation in DNA methylation, and thus are better able to stratify normal tissues, cellular phenotypes, or patient outcomes [50–52]. While hypermethylation of normally unmethylated CpG island containing promoters is a well-described mechanism for the inactivation of tumor suppressor genes in cancer, recent studies underscore the contribution of altered methylation at enhancer elements as an important contributor to the aberrant gene expression programs that define human cancers [51, 53, 54]. Taken together, our data suggest that the stable imprint of a prior high-LET radiation exposure is reflected in the DNA methylation pattern, and may prove useful as a biomarker for long-term, individual cancer risk.

Figures

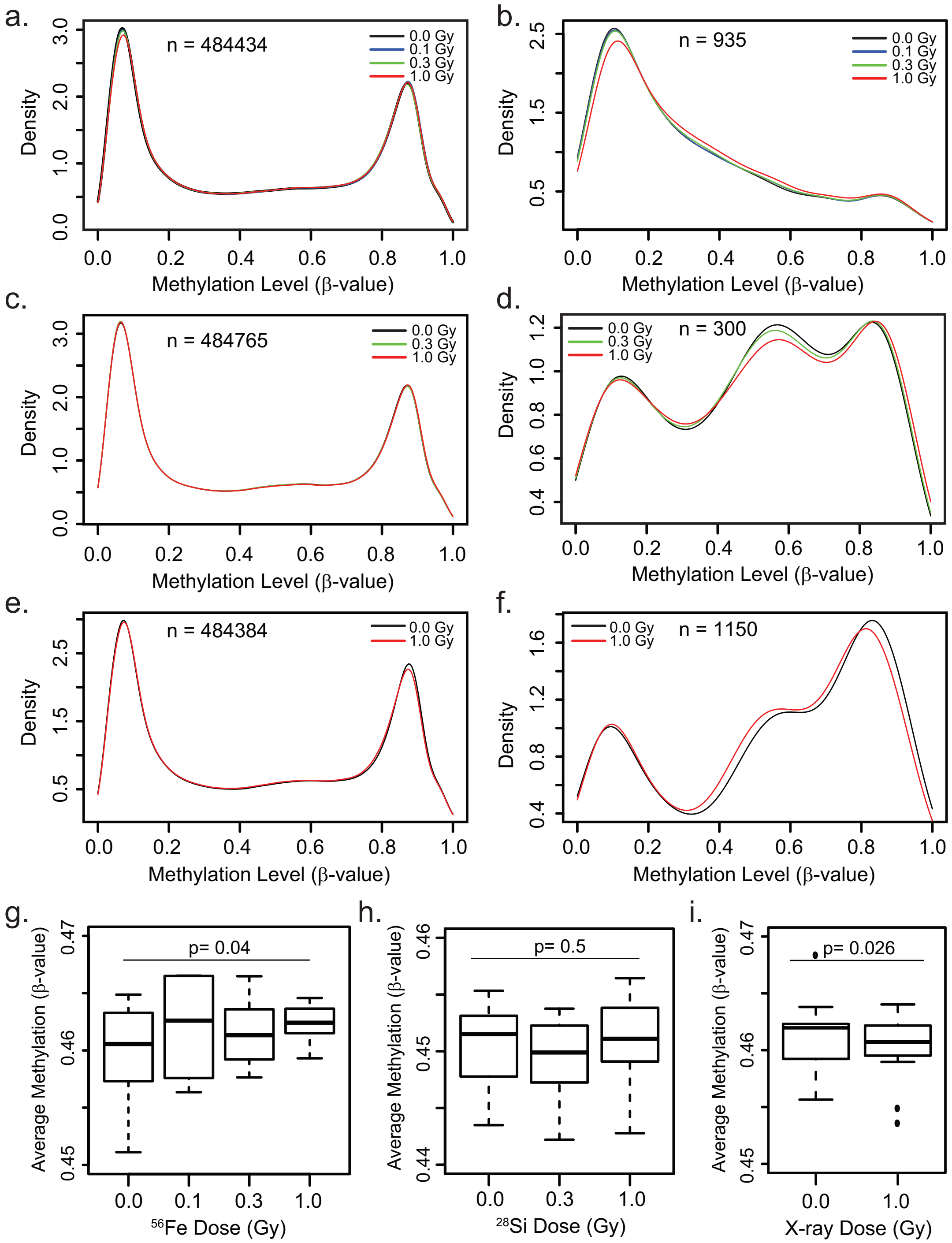


Figure 3-1. Global impact of High- vs. Low- LET radiation on DNA methylation. a-f) Density plots showing the impact of the indicated dose of each radiation source on the distribution of DNA methylation (β values) across all sites (**a**,**c**,**e**) or the subset of

sites whose methylation status was found to be significantly associated with dose of ^{56}Fe, ^{28}Si or X ray (**b**, **d**, **f**). **g-i**) Box plot distribution of the average methylation level across all >484,000 CpG sites for probes passing QC. Line represents the median, boxes the first and third quartiles, whiskers represent the interquantile distance. Note the trend towards hypermethylation with increasing ^{56}Fe ion dose (p=0.04, Mann-Whitney) and towards hypomethylation with X ray dose (p=0.026, Mann-Whitney). No significant directional trend was observed with ^{28}Si exposure.

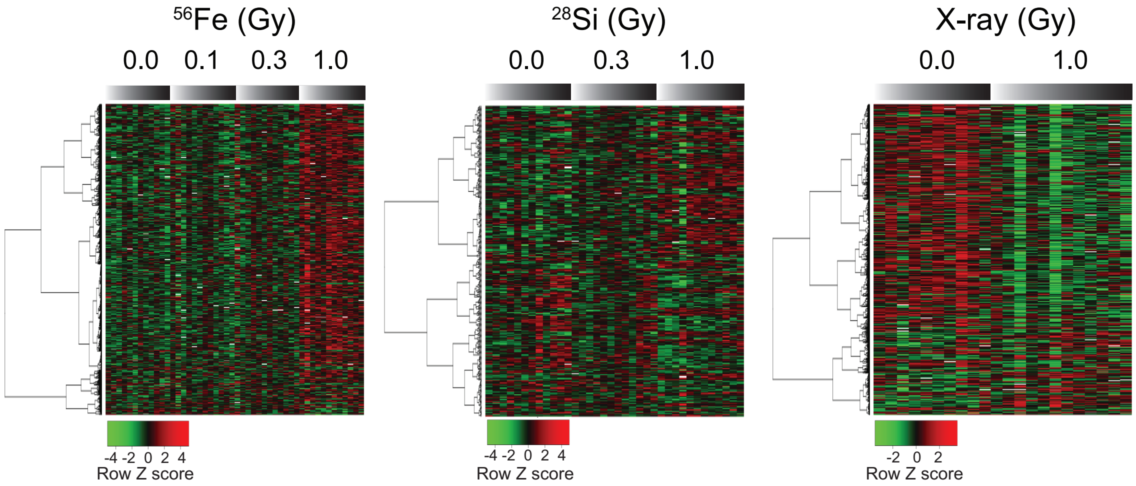


Figure 3-2. Differential effects of High- (^{56}Fe, ^{28}Si) or low-LET radiation dose on the methylation status of individual CpG sites.

A linear mixed effects model was used to identify DNA methylation changes significantly associated with dose, source, or time-after-exposure. This analysis identified 934 CpG sites whose methylation status was moderately ($p<0.001$) associated with ^{56}Fe dose (849 hyper; 86 hypo); 300 CpG sites associated with ^{28}Si dose (158 hyper, 142 hypo) and 1150 CpG sites associated with X ray dose (252 hyper; 898 hypo). Heatmap showing the methylation status (low, green to high, red) of the ^{56}Fe, ^{28}Si or X ray significant CpG sites (rows) methylation across all samples analyzed (columns). Columns are grouped according to dose, and arranged by increasing time after exposure (Gray bar).

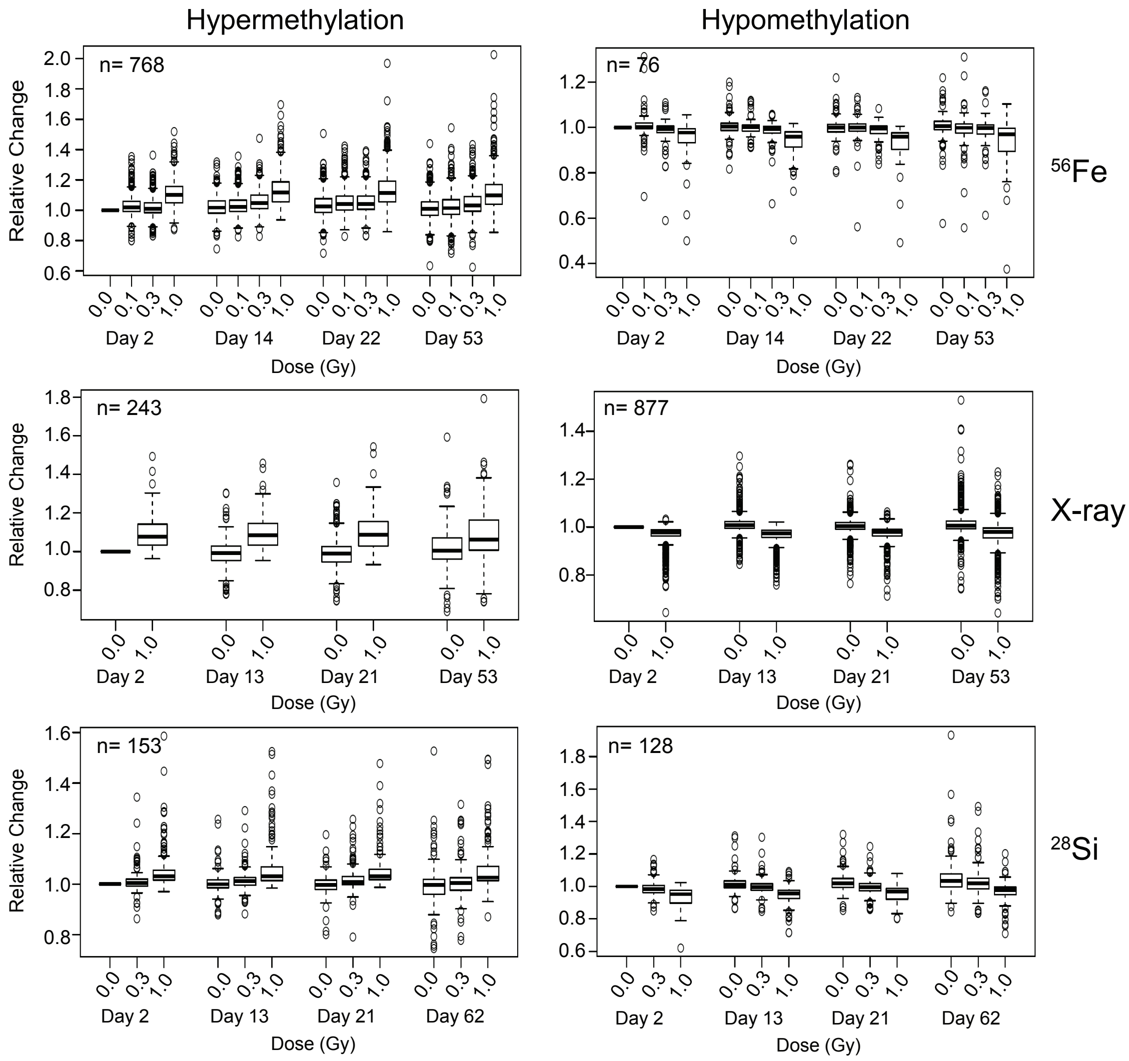


Figure 3-3. Fate of Radiation-induced changes in DNA methylation over time.

CpG sites exhibiting a change in methylation level significantly associated with radiation dose were normalized to their individual initial methylation levels as extrapolated from the 48h, unexposed cultures. Shown is the distribution of the change in methylation of CpG sites undergoing hyper- or hypo-methylation in response to the indicated radiation source relative to the internal control (unexposed) cultures at 48h. Line represents the median, boxes the first and third quartile, and whiskers extend to maximum value that is 1.5 times the interquartile range. For clarity, CpG sites whose methylation level was also independently associated with time-after-culture were excluded. Note that radiation-induced methylation changes occur early (within 48h) and largely persist over time.

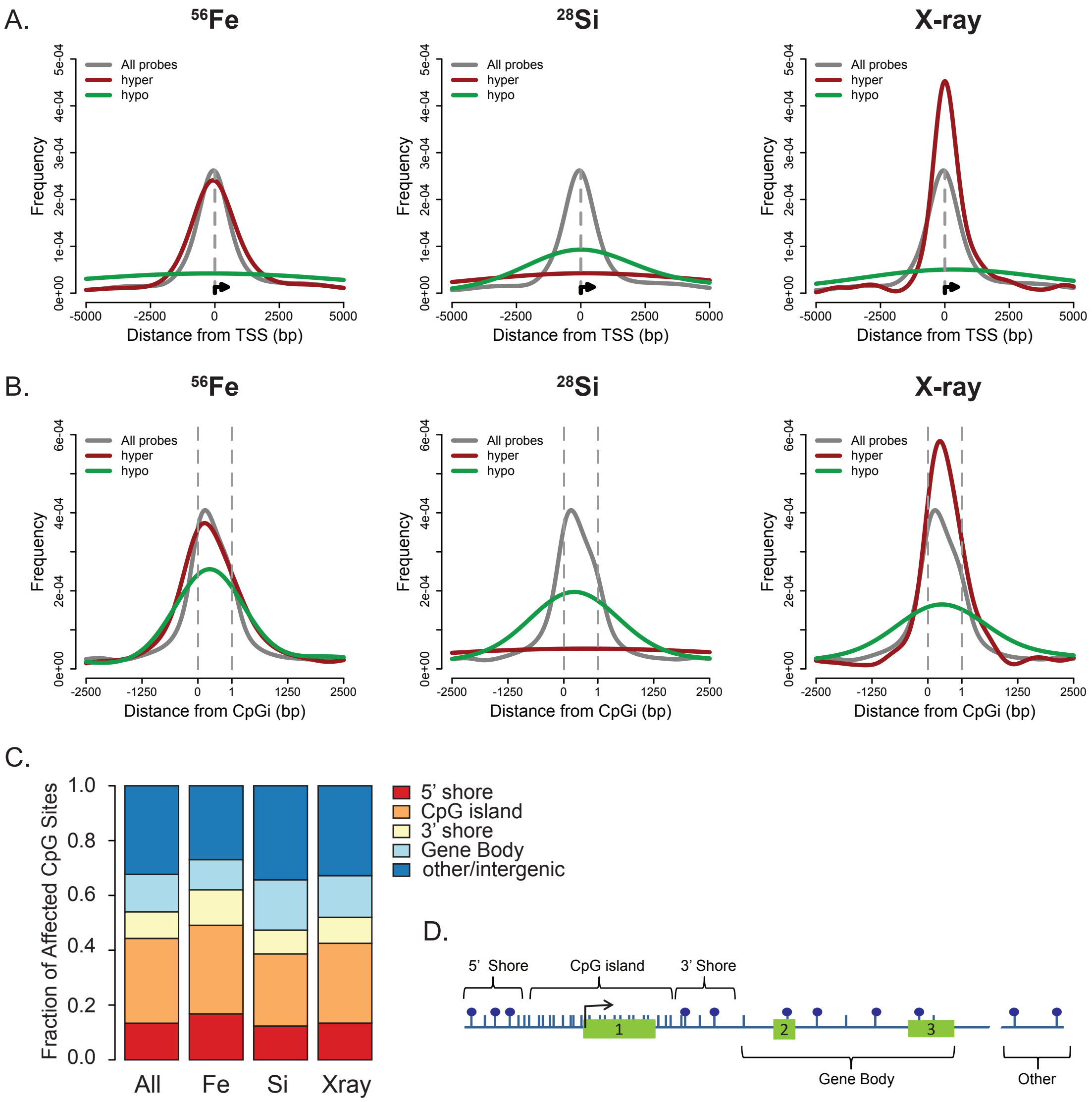


Figure 3-4. Genomic location of CpG sites significantly associated with radiation dose.

a) Average distance of all CpGs on the array (gray), or the subset that underwent hyper (red) or hypo (green) methylation in response to increasing ^{56}Fe, ^{28}Si or X ray dose relative to the transcription start site (TSS) of the nearest gene oriented to the direction of transcription. **b)** Average distance of all CpGs on the array (gray), or the subset that underwent hyper (red) or hypo (green) methylation in response to increasing ^{56}Fe, ^{28}Si or X ray dose relative to the nearest CpG island. The distribution within the CpG island is scaled to size (dotted gray lines), and includes a fixed distance of +/-2.5 kb in either direction from the CpG island edge. **c)** Fraction of ^{56}Fe, ^{28}Si or X ray affected CpG sites that lie within the indicated gene compartment relative to that of All CpG sites interrogated on the array. CpG sites were annotated to the nearest CpG island associated RefSeq gene. CpG islands defined by UCSC criteria, 5' and 3' shores are 2,000 bp from the 5' and 3' CpG island edge. Gene bodies were considered the region from the 3' edge of the CpG island +2 kb, to the transcription end site (TES). CpG sites not overlapping one of these features were considered to be intergenic/other. **d)** A schematic of the genomic compartments described in *C*. Shown is a hypothetical gene (exons-green boxes) for which the TSS (black arrow) is embedded in a CpG island promoter. Blue ticks represent CpG sites, blue balls as methylated CpG sites. The TES would be the end of exon 3.

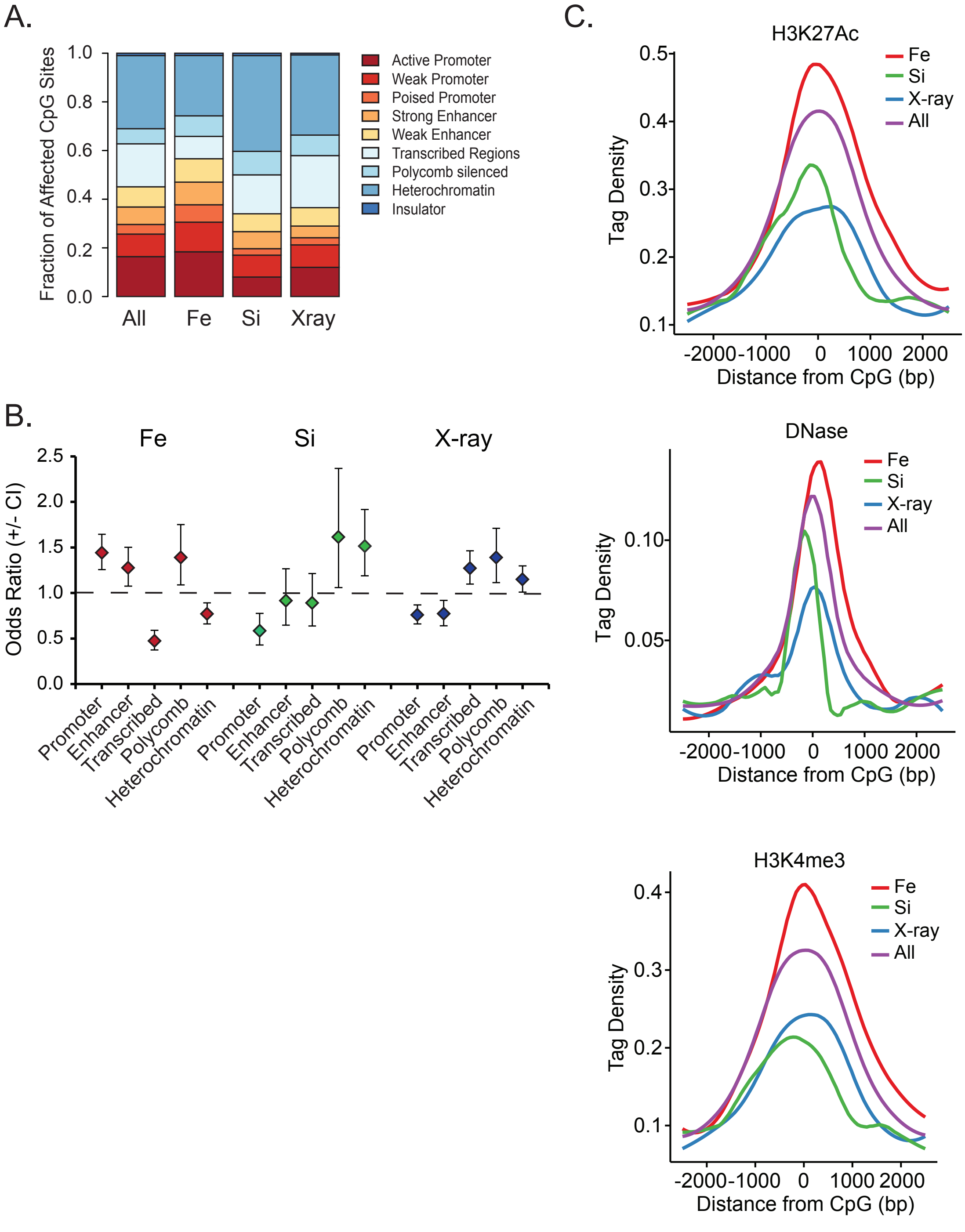
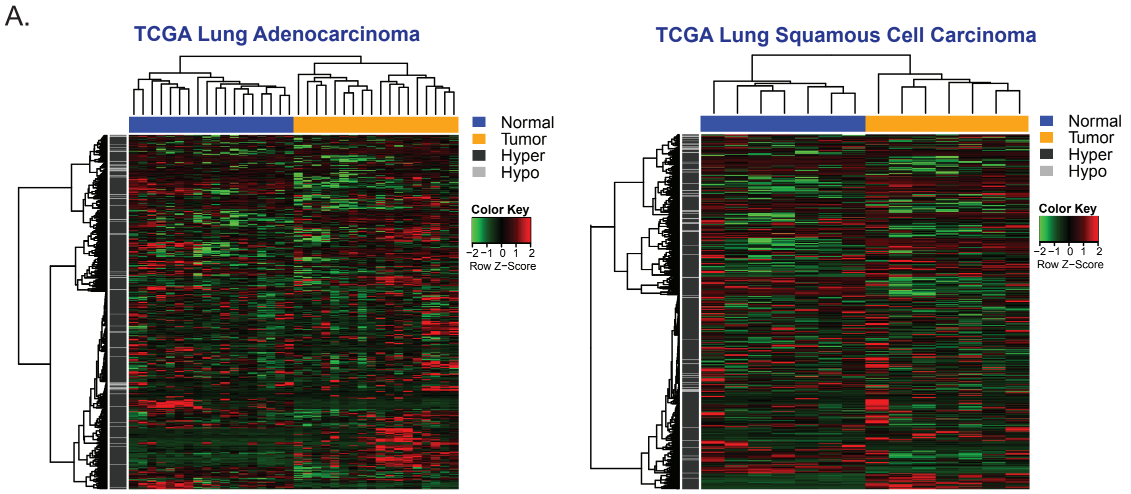


Figure 3-5. HZE-particles of distinct LETs affect methylation of CpG sites in different genomic chromatin compartments. CpG sites were annotated to a chromatin-based functional genomics annotation, ChromHMM, established by Ernst et al. [27] using 14 different chromatin features from ENCODE data from human epithelial cells

(HMEC). Shown is the fraction **(a)** and relative enrichment **(b)** of ^{56}Fe, ^{28}Si or X ray affected CpG sites that overlap in the indicated compartment, relative to that of ‘All’ CpG sites on the array. Data represent the odds ratios determined by Fishers exact +/- the 95th confidence interval. (**C**) Normalized average tag densities of H3K27 acetylation ChIP-seq (*Top*), DNaseI-seq (*Middle*) or H3K4me3 ChIP-seq (*Bottom*) surrounding all assayed CpGs (All), or the subset of CpGs whose change in methylation was significantly associated with ^{56}Fe, ^{28}Si or X ray dose. Data are derived from ENCODE CHIP-seq and DNAse-I seq data from A549 lung cancer cells [25]. Note the over-representation of H3K4me3, H3K27ac, and DNaseI accessibility at ^{56}Fe-affected sites.

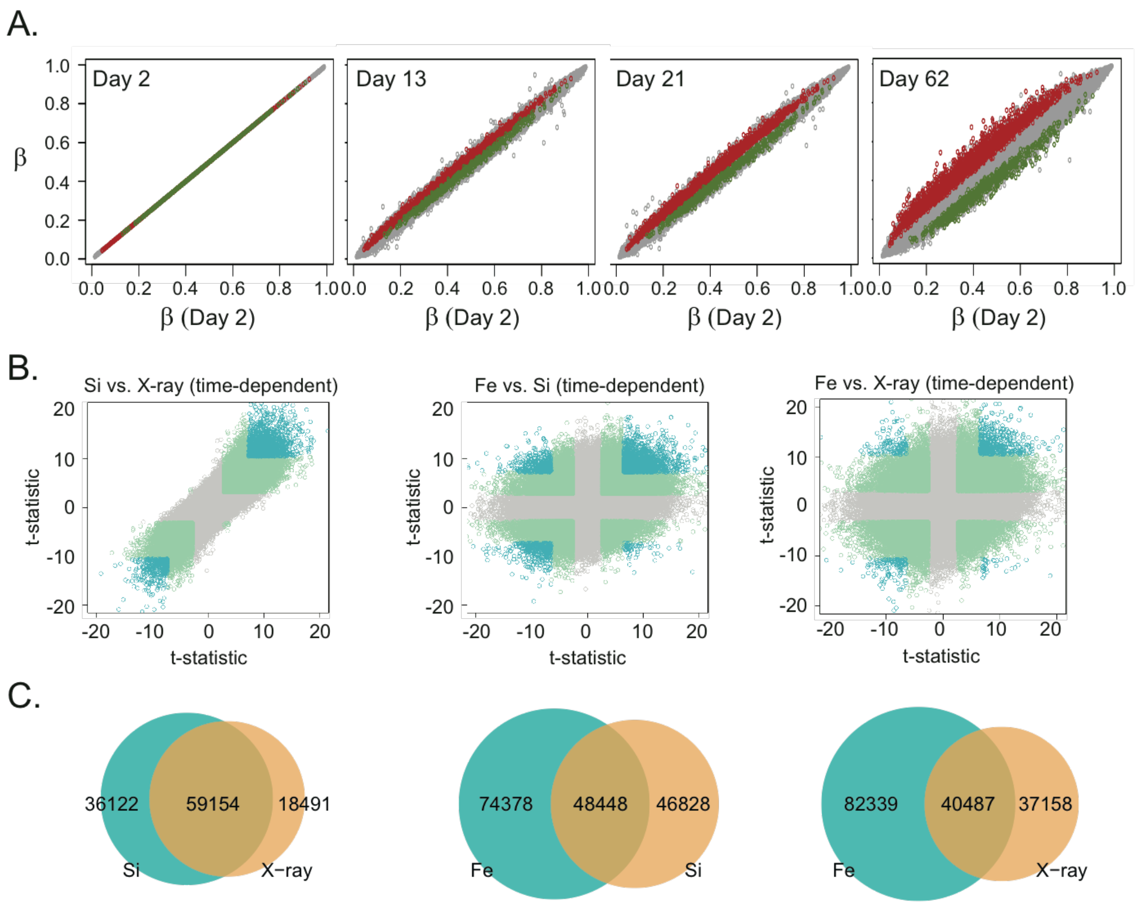


B.

	Tumor-Normal samples (n)	^{56}Fe	^{28}Si	X-ray
LuAC	36 (18 pairs)	*CpG sites=777* p-value=1.456e-08	*CpG Sites=236* p-value=0.4669	*CpG Sites=934* p-value=0.4669
LuSC	14 (7 pairs)	*CpG sites=782* p-value=0.001341	*CpG sites=237* p-value=0.1927	*CpG Sites=938* p-value=0.1927

Figure 3-6. The ^{56}Fe-specific methylation ‘signature’ discriminates lung tumor from normal tissue in primary tissue samples.

a) DNA methylation status of the CpG sites significantly associated with Fe dose (n=777) in normal bronchial epithelial cells was extracted for 25 lung tumor-normal pairs (18 adenocarcinomas, 7 squamous cell carcinomas) available in the TCGA project and used in unsupervised hierarchical cluster analysis (complete linkage, Manhattan distance metric). **b)** An equivalent number of CpGs were chosen at random and used to group tissue samples using the same approach, and the process was repeated 1,000 times to estimate significance. The ^{56}Fe-sensitive CpGs outperformed any random set by several orders of magnitude. The methylation status of the ^{28}Si or the X ray affected sites had no significant association with tumor-specific differences in methylation (see Methods).



Supplemental Figure S3-1. Time-dependent methylation drift

a) Comparison of the average methylation (β-value) of each CpG site in the X ray-exposed cohort at the indicated time-after-exposure relative to that on Day 2. Red indicates those CpGs that are significantly hypermethylated with time (n=2,294); green are those sites significantly hypomethylated with time (n=647; p<1e-7). **b)** A linear mixed effects model was applied to identify DNA methylation changes significantly associated with time-after-exposure for each exposure type. Shown are scatter plots comparing the significance and direction of change (t-statistics) for CpG sites at which methylation changed significantly with time-after-exposure in the ^{28}Si exposed series, the ^{56}Fe exposed series, and the X ray exposed series. A positive t-statistic indicates a gain in methylation (hypermethylated) and a negative t-statistic indicates a loss of methylation (hypomethylation). Light green indicates sites reaching an FDR<0.05 (Benjamini-Hochberg) and blue are those reaching Holm significance (p<1e-7). **c)** Venn diagrams comparing the overlap of CpG sites whose methylation change was significantly associated with time in culture [FDR<0.05 (Benjamini-Hochberg)] in the ^{28}Si exposed series, the ^{56}Fe exposed series, and the X ray exposed series. **d)** Average methylation level (β) over time among CpG sites negatively (left) or positively (right) associated with time stratified by ^{56}Fe-ion dose. Analysis was restricted to those CpG sites found to be independently associated with both dose and time in the ^{56}Fe-ion exposed series (n=91).

References

1. Held KD. Effects of low fluences of radiations found in space on cellular systems. Int J Radiat Biol. 2009;85:379–90.
2. Hu W, Pei H, Li H, Ding N, He J, Wang J, et al. Effects of shielding on the induction of 53BP1 foci and micronuclei after Fe ion exposures. J Radiat Res (Tokyo). 2014;55:10–6.
3. Lebel EA, Rusek A, Sivertz MB, Yip K, Thompson KH, Tafrov ST. Analyses of the Secondary Particle Radiation and the DNA Damage It Causes to Human Keratinocytes. J Radiat Res (Tokyo). 2011;52:685–93.
4. Mukherjee B, Camacho CV, Tomimatsu N, Miller J, Burma S. Modulation of the DNA-damage response to HZE particles by shielding. DNA Repair. 2008;7:1717–30.
5. Durante M, Cucinotta FA. Heavy ion carcinogenesis and human space exploration. Nat Rev Cancer. 2008;8:465–72.
6. Cucinotta FA, Durante M. Cancer risk from exposure to galactic cosmic rays: implications for space exploration by human beings. Lancet Oncol. 2006;7:431–5.
7. Plante I, Ponomarev AL, Cucinotta FA. Calculation of the energy deposition in nanovolumes by protons and HZE particles: geometric patterns of initial distributions of DNA repair foci. Phys Med Biol. 2013;58:6393–405.
8. Cucinotta FA, Cacao E. Non-Targeted Effects Models Predict Significantly Higher Mars Mission Cancer Risk than Targeted Effects Models. Sci Rep. 2017;7:1832.
9. Shinoto M, Ebner DK, Yamada S. Particle Radiation Therapy for Gastrointestinal Cancers. Curr Oncol Rep. 2016;18:17.

10. Ding L-H, Park S, Peyton M, Girard L, Xie Y, Minna JD, et al. Distinct transcriptome profiles identified in normal human bronchial epithelial cells after exposure to γ-rays and different elemental particles of high Z and energy. BMC Genomics. 2013;14:372.

11. Huidobro C, Fernandez AF, Fraga MF. Aging epigenetics: causes and consequences. Mol Aspects Med. 2013;34:765–81.

12. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.

13. Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, et al. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature. 2008;455:822–5.

14. Filkowski JN, Ilnytskyy Y, Tamminga J, Koturbash I, Golubov A, Bagnyukova T, et al. Hypomethylation and genome instability in the germline of exposed parents and their progeny is associated with altered miRNA expression. Carcinogenesis. 2010;31:1110–5.

15. Goetz W, Morgan MNM, Baulch JE. The effect of radiation quality on genomic DNA methylation profiles in irradiated human cell lines. Radiat Res. 2011;175:575–87.

16. Ilnytskyy Y, Koturbash I, Kovalchuk O. Radiation-induced bystander effects in vivo are epigenetically regulated in a tissue-specific manner. Environ Mol Mutagen. 2009;50:105–113.

17. Pogribny I, Raiche J, Slovack M, Kovalchuk O. Dose-dependence, sex- and tissue-specificity, and persistence of radiation-induced genomic DNA methylation changes. Biochem Biophys Res Commun. 2004;320:1253–61.

18. Antwih DA, Gabbara KM, Lancaster WD, Ruden DM, Zielske SP. Radiation-induced epigenetic DNA methylation modification of radiation-response pathways. Epigenetics. 2013;8:839–48.

19. Bae J-H, Kim J-G, Heo K, Yang K, Kim T-O, Yi JM. Identification of radiation-induced aberrant hypomethylation in colon cancer. BMC Genomics. 2015;16:56.

20. Ramirez RD, Sheridan S, Girard L, Sato M, Kim Y, Pollack J, et al. Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins. Cancer Res. 2004;64:9027–34.

21. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinforma Oxf Engl. 2012;28:1280–1.

22. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer New York; 2005. p. 397–420. http://link.springer.com/chapter/10.1007/0-387-29362-0_23. Accessed 20 Jul 2013.

23. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat. 1979;6:65–70.

24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.

25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

26. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013;9:e1003118.

27. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.

28. Lima F, Ding D, Goetz W, Yang AJ, Baulch JE. High LET 56Fe ion irradiation induces tissue-specific changes in DNA methylation in the mouse. Environ Mol Mutagen. 2014;55:266–77.

29. Nzabarushimana E, Miousse IR, Shao L, Chang J, Allen AR, Turner J, et al. Long-term epigenetic effects of exposure to low doses of 56Fe in the mouse lung. J Radiat Res (Tokyo). 2014;55:823–8.

30. Rithidech KN, Jangiam W, Tungjai M, Gordon C, Honikel L, Whorton EB. Induction of Chronic Inflammation and Altered Levels of DNA Hydroxymethylation in Somatic and Germinal Tissues of CBA/CaJ Mice Exposed to (48)Ti Ions. Front Oncol. 2016;6:155.

31. Cardis E, Vrijheid M, Blettner M, Gilbert E, Hakama M, Hill C, et al. The 15-Country Collaborative Study of Cancer Risk among Radiation Workers in the Nuclear Industry: estimates of radiation-related cancer risks. Radiat Res. 2007;167:396–416.

32. Preston DL, Ron E, Tokuoka S, Funamoto S, Nishi N, Soda M, et al. Solid cancer incidence in atomic bomb survivors: 1958-1998. Radiat Res. 2007;168:1–64.

33. Cucinotta FA. Space Radiation Risks for Astronauts on Multiple International Space Station Missions. PLoS ONE. 2014;9:e96099.

34. Suman S, Kumar S, Moon B-H, Strawn SJ, Thakor H, Fan Z, et al. Relative Biological Effectiveness of Energetic Heavy Ions for Intestinal Tumorigenesis Shows Male Preponderance and Radiation Type and Energy Dependence in APC(1638N/+) Mice. Int J Radiat Oncol Biol Phys. 2016;95:131–8.

35. Trani D, Datta K, Doiron K, Kallakury B, Fornace AJ. Enhanced Intestinal Tumor Multiplicity and Grade in vivo after HZE Exposure: Mouse Models for Space Radiation Risk Estimates. Radiat Environ Biophys. 2010;49:389–96.

36. Wang X, Farris III AB, Wang P, Zhang X, Wang H, Wang Y. Relative Effectiveness at 1 Gy after Acute and Fractionated Exposures of Heavy Ions with Different Linear Energy Transfer for Lung Tumorigenesis. Radiat Res. 2015;183:233–9.

37. Weil MM, Bedford JS, Bielefeldt-Ohmann H, Ray FA, Genik PC, Ehrhart EJ, et al. Incidence of acute myeloid leukemia and hepatocellular carcinoma in mice irradiated with 1 GeV/nucleon (56)Fe ions. Radiat Res. 2009;172:213–9.

38. Cucinotta FA, Chappell LJ. Updates to astronaut radiation limits: radiation risks for never-smokers. Radiat Res. 2011;176:102–14.

39. Ding N, Bonham EM, Hannon BE, Amick TR, Baylin SB, O’Hagan HM. Mismatch repair proteins recruit DNA methyltransferase 1 to sites of oxidative DNA damage. J Mol Cell Biol. 2016;8:244–54.

40. O’Hagan HM. Chromatin modifications during repair of environmental exposure-induced DNA damage: a potential mechanism for stable epigenetic alterations. Environ Mol Mutagen. 2014;55:278–91.

41. O'Hagan HM, Mohammad HP, Baylin SB. Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island. PLoS Genet. 2008;4:e1000155.

42. O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, et al. Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. Cancer Cell. 2011;20:606–19.

43. Acharya MM, Baddour AAD, Kawashita T, Allen BD, Syage AR, Nguyen TH, et al. Epigenetic determinants of space radiation-induced cognitive dysfunction. Sci Rep. 2017;7. doi:10.1038/srep42885.

44. Lorat Y, Timm S, Jakob B, Taucher-Scholz G, Rübe CE. Clustered double-strand breaks in heterochromatin perturb DNA repair after high linear energy transfer irradiation. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2016;121:154–61.

45. Bielefeldt-Ohmann H, Genik PC, Fallgren CM, Ullrich RL, Weil MM. Animal studies of charged particle-induced carcinogenesis. Health Phys. 2012;103:568–76.

46. Datta K, Suman S, Kallakury BVS, Fornace AJ. Heavy ion radiation exposure triggered higher intestinal tumor frequency and greater β-catenin activation than γ radiation in APC(Min/+) mice. PloS One. 2013;8:e59295.

47. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41:178–86.

48. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

49. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.

50. Aran D, Hellman A. Unmasking risk loci: DNA methylation illuminates the biology of cancer predisposition: analyzing DNA methylation of transcriptional enhancers reveals missed regulatory links between cancer risk loci and genes. BioEssays News Rev Mol Cell Dev Biol. 2014;36:184–90.

51. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14:1–14.

52. Wiench M, John S, Baek S, Johnson TA, Sung M-H, Escobar T, et al. DNA methylation status predicts cell type-specific enhancer activity. EMBO J. 2011;30:3028–39.

53. Bell RE, Golan T, Sheinboim D, Malcov H, Amar D, Salamon A, et al. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. Genome Res. 2016;26:601–11.

54. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 2014.

Chapter IV. An integrated -omics analysis of the epigenetic landscape of gene expression in human blood cells

Elizabeth M. Kennedy, George N. Goehring, Michael H. Nichols, Chloe Robins, Divya Mehta, Torsten Klengel, Eleazar Eskin, Alicia K. Smith, Karen N. Conneely

Adapted from a manuscript currently submitted to *PLOS Genetics*

Introduction

DNA methylation occurs at CG dinucleotides (CpGs) and is an essential epigenetic mechanism for many organisms. Regions of CpG-rich sequences, termed CpG islands, are found throughout the human genome. These CpG islands overlap with promoter regions or transcription factor binding sites for approximately half of mammalian genes, including nearly all housekeeping genes [1]. Canonically, methylation in promoter CpG islands inhibits the initiation of gene transcription [2]. Through modulation of gene transcription and expression, epigenetic modifications allow for morphologically distinct cell types to form from a single genome [3, 4]. Epigenome-wide association studies (EWAS) have also linked certain DNA methylation patterns to environmental factors, aging, and disease [5–14].

Unfortunately, despite a growing number of EWAS, we are still far from understanding how epigenetic changes contribute to the onset of complex diseases [2, 15]. EWAS often return large sets of marginally significant or near-significant results, many of which lie outside of defined genomic regions (i.e. genes) [16, 17]. Inferring a functional consequence of such results is difficult because our understanding of the role of methylation in gene expression is incomplete. This is especially true for EWAS hits outside promoters, as the role of DNA methylation in these regions is not fully defined [2].

Recent studies have set out to clarify the role of DNA methylation in gene expression by investigating associations between gene expression and the methylation of nearby CpGs. CpGs with methylation changes that associate with expression changes are called expression-associated CpGs, or eCpGs. The results of these studies suggest that gene transcription can be influenced by DNA methylation at CpGs that are distal (>50 kb or on a different chromosome) to the gene promoter [18–22]. Additionally, many of these studies report that changes to CpG methylation in enhancers may be central to epigenetic gene regulation. However, most of these studies tested only for eCpGs within a limited distance from each gene [18, 21–23], with few seeking to identify genome-wide eCpGs for each gene [19, 20]. In this study, we define genome-wide epigenetic signatures for more than 13k transcripts, based on methylation at over 420k individual CpGs in two human studies. We find evidence that CpG methylation changes associate with gene expression at great distances throughout the genome. Our results broaden the understanding of epigenetics and gene regulation and have the potential to provide critical biological insight for new and existing EWAS.

Materials and methods

Data preprocessing and QC. The Grady Trauma Project (GTP) is a cross-sectional study of stress-related outcomes. Participants were recruited from the waiting rooms of Grady Memorial Hospital's General Practice or Obstetrics and Gynecology departments in Atlanta, GA. Participants are from an inner-city population with higher than average rates of trauma exposure, but are representative of this population as they are not specifically ascertained for presence of disease or trauma. Genome-wide DNA methylation and gene expression measurements were generated for 333 human blood samples. GTP participants included in this study range between 18 and 78 years old, are 76% female and all are African-American.

The Multi-Ethnic Study of Atherosclerosis (MESA) is a study designed to examine cardiovascular disease. The MESA Epigenomics and Transcriptomics Study specifically investigates the association between CpG methylation and gene expression in purified human monocytes collected from the MESA population. For this study, 1,202 participants were chosen randomly from samples collected between April 2010 and February 2012 from MESA field centers in Baltimore, MD; Forsyth County, NC; New York, NY; and St Paul, MN. Participants range in age from 55 – 94 years old, are 51% female, and self identified as Caucasian (47%), African American (21%), or Hispanic (32%) [23].

For both GTP and MESA, methylation data for >480K individual CpGs were generated from the Infinium HumanMethylation450 BeadChip (Illumina, San Diego, CA), and RNA transcript levels for >25,000 annotated genes were quantified via Illumina HumanHT-12 v3.0 and v4.0 Expression BeadChip. We have provided a detailed description of both datasets, including sample information, data processing, QC, and normalization, in the supplemental methods. We excluded CpGs and transcripts that did not pass QC, were on the X or Y chromosomes or were poor quality. After QC, 13,933 expression probes (transcripts) and 483,399 CpG probes (CpGs) remained for GTP, and 19,445 transcripts and 422,016 CpGs remained for MESA.

Association analysis. To model the associations between gene expression and CpG methylation at specific sites while adjusting for global expression and methylation differences between individuals, we used a linear mixed model framework developed to account for inter-individual correlation structure in expression data due to unknown confounders (inter-sample correlation emended or ICE; [24]. For all transcripts and CpGs in each study, we regressed log expression signals for one transcript on methylation β-values for a single CpG, while controlling for fixed effects (age and sex for GTP and age and composite age/gender/study-site for MESA) and unknown random effect covariates using ICE (equation 1). We implemented this framework in the python program pyLMM (http://genetics.cs.ucla.edu/pylmm/) to test for association between methylation at CpG *j* and the expression level of transcript *k*, by fitting the model:

$$y_{\mathrm{k}} = \mu_{\mathrm{k}} + M_{\mathrm{j}}\,\alpha_{\mathrm{jk}} + x\,\beta_{\mathrm{jk}} + u_{\mathrm{k}} + \epsilon_{\mathrm{jk}} \qquad (1)$$

Letting *n* be the number of individuals, y_{k} is a vector of log expression levels at gene *k* with length *n*, μ_{k} is a size *n* vector denoting the mean of log expression levels over *n* individuals, M_{j} is a size *n* vector of methylation proportions at CpG *j*, *x* is an *n*×2 matrix of covariates (age and sex), $u_{\mathrm{k}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{g}}\mathrm{H})$ is a multivariate normally distributed term representing effects due to other unmeasured confounders such as cellular heterogeneity, and $\epsilon_{\mathrm{jk}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{e}}\mathrm{I})$ are residual errors. I is an *n*×*n* identity matrix and H is the *n*×*n* intersample correlation matrix, described below.

Intersample correlation matrix. The global intersample correlation matrix H is estimated from the expression data. Let Y be an *m*×*n* expression matrix for *m* genes and *n* individuals. Then let Z be an *m*×*n* matrix where each element from the $\underline{k}^{\mathrm{th}}$ transcript and l^{th} individual $Z_{\mathrm{kl}}= (y_{\mathrm{kl}}-\mu_{\mathrm{k}})/\sigma_k$; μ_{k} is the mean and σ_k is the standard deviation of log expression values of the k^{th} transcripts. The estimated intersample correlation matrix $\hat{\mathrm{H}}$, is defined as the covariance of Z, and is in equation (1) to correct for unmeasured confounding factors.

Analysis of results. In the association analysis, we analyzed all combinations of transcripts and CpGs, for a total of 6.6 billion comparisons for GTP and 8.2 billion comparisons for MESA. For each transcript, pyLMM generated summary statistics for the association of all CpGs. Based on these statistics, genomic inflation factors (GIF) were calculated as median (T-statistic)^2/0.4549 for each transcript. We removed transcripts with a GIF >2 from further analysis. We also removed CpG-transcript pairs in which the associated transcript was annotated as bad quality or as having no matching sequence in the genome [25].

A re-annotation of the Illumina HumanHT-12 v3.0 and v4.0 Expression BeadChip arrays by Barbosa-Morais and others (2010) indicates that many probes have the potential to anneal to multiple regions in the genome, by sequence homology (determined via BLAST and BLAT searches; [25]. This non-specific binding could potentially lead to an inaccurate picture of eCpG-transcript associations, especially when the potential binding locations for an expression probe are located on multiple chromosomes. To avoid this issue, we allowed each expression probe to have multiple locations, based on the new annotation. Using the refseq and ensembl databases [26, 27], we assigned each expression probe location to a gene by overlap with an exon. We chose the location of the expression probe for each eCpG-transcript association, prioritizing expression probe locations that were closer in proximity to the eCpG, could be annotated to a gene and were listed by Barbosa-Morais as the primary>secondary> other genomic match (see supplemental methods).

To establish a similar cutoff for significance across GTP and MESA, we considered CpG-transcript pairs with $p<10^{-5}$ as suggestive and $p<10^{-11}$ as significant. This value corresponds to Bonferroni adjustment for 5 billion independent tests, so is quite conservative given the high levels of correlation between tests. We defined CpGs that significantly associate with transcript expression as eCpGs.

We classified eCpGs, broadly, as cis (within 50kb of associated probe), distal (greater than 50kb from associated gene, but on the same chromosome) or trans (on a different chromosome from the associated gene). Within those broad categories, we established the following detailed classifications to describe each eCpG-transcript pair with respect to the gene the associated transcript is annotated to, as well as other nearby genes (by average refseq and ensembl gene locations (see transcript annotation in supplemental methods)): **trans** (the eCpG was on a different chromosome than the transcript), **distal** (the eCpG was >50kb from the transcript, but on the same chromosome), **in gene body** (the eCpG was >2,500bp downstream of the associated gene's TSS and upstream of the associated gene's TES), **near promoter** (the eCpG was within 2,500bp upstream or downstream of the associated gene's TSS), **closest upstream gene** (the TES of the associated gene was closer to the eCpG than the next closest gene), **closest downstream gene** (the eCpG was not within 2,500bp of the associated gene, but the TSS of the associated gene was closer to the eCpG than the next closest gene), **closer 5'** (the eCpG was farther from the associated gene's TSS than from another gene's TSS on the opposite side of the eCpG), **closer 3'** (the eCpG was closer to the TES of another gene than the associated gene than to either the TSS or TES of the associated gene), **gene between** (there was another gene's TSS between the eCpG and the associated gene's TSS), **eCpG in different gene** (the eCpG was not near the promoter of the associated gene and was between the TSS and TES of another gene), **multiple closer/between** (the eCpG-transcript pair falls into multiple of the aforementioned cis categories; Figure 4-1).

Between study corroboration. Next we sought to find out how often GTP eCpG-transcript pairs were consistent in MESA results among the cis, distal and trans categories. To compare results between studies, we found eCpG-transcript pairs in the GTP results that were consistent in the MESA results by CpG ID, expression probe ID, expression probe location, and direction of correlation.

To compare the number of eCpG-transcript pairs found consistent across studies within the distal and trans categories to the number achieved by random chance, we re-analyzed the results 10,000 times. For each permutation, we randomly shuffled the expression probe IDs within each study and category.

Within study corroboration. For each eCpG found in both GTP and MESA, we interrogated neighboring eCpGs within five windows extending 100, 500, 1,000, 1,500 and 2,000 bp to each side of the query eCpG. For each window, we compared the genes associated (see transcript annotation in supplemental methods) with the query eCpG to the genes associated with the neighboring eCpGs. We computed the percentage of eCpGs sharing an associated gene with a neighboring eCpG as the number of eCpGs that share at least one associated gene with at least one neighboring eCpG divided by the total number of eCpGs within the window. We then computed the percentage of eCpGs sharing at least one associated gene with the same direction of correlation with at least one neighboring eCpG. Lastly we computed the percentage of eCpGs at which all neighboring eCpGs shared both genes and direction of correlation with the query eCpG. This analysis was conducted for all eCpGs and then separately for trans eCpGs.

Functional analysis of eCpGs. We downloaded the following datasets from the UCSC table browser for GRCh37/hg19 [28]:

1) CpG Islands
2) Broad ChromHMM for GM12878 [29]

3) Transcription factor ChIP V3 (transcription factor binding sites)

We functionally annotated eCpGs based on overlap of the CpG location with the intervals provided by UCSC for the features listed above. Additionally, CpG island shores were defined as regions extending 1.5kb out from CpG islands and CpG island shelves were defined as regions extending 1.5kb out from shores. Intervals for all 15 genomic states provided with the ChromHMM dataset were utilized in this annotation. We assessed these annotations, using Fisher's exact tests, in two different ways. First we considered all CpGs tested for each study. Each CpG was only represented once (for each study) and was tested for enrichment in a functional category (e.g. CpG island, ChromHMM category) and significant eCpG status (i.e. significant vs not significant). Second, among only significant eCpG-transcript pairs, eCpG-transcript classifications (e.g. "in gene", "closest upstream gene"; described above and in Figure 4-1) were tested for enrichment among the various functional categories (e.g. CpG island, ChromHMM category). Because many CpGs associated with multiple transcripts, and vice versa, CpGs or transcripts could fall into more than one category and be present more than once in the test. However, each unique CpG-transcript pair falls into a single category and is present only once in the test.

Genomic interaction distance decay. In this analysis we used the GM12878 Hi-C dataset of Rao et. al 2014 accessed from GSE63525 [30]. The average number of interactions between each 1 kb bin was taken from the expected values for unnormalized interaction counts on Chromosome 1 at 1 kb. We next calculated the distance (rounded to the nearest 1 kb) between eCpGs and the transcription start sites (TSS) annotated to their associated expression probes for each of the 4,799 significant cis and distal eCpG-transcript pairs.

To plot the distance decay of both datasets, the logarithmic space between 10^3 and 10^8 was divided into 30 equally spaced bins. Means and 0.95 confidence intervals were computed for each bin using the .regplot function in the python package, Seaborn (DOI 10.5281/zenodo.592845).

Gene ontology analyses. We used the R library GOstats [31] to assess enrichment of molecular function gene ontology terms among eCpGs. eCpGs that associated with a transcript with p-values $<10^{-5}$ were included in the analysis. We applied the hypergeometric test to calculate odds ratios and p-values, and estimated the false discovery rate by the Benjamini & Hochberg method [32]. For this analysis, eCpGs that did not fall within a gene were assigned the Entrez gene ID of the gene with the closest downstream TSS. We assessed eCpGs in the following scenarios: all eCpGs, cis and distal eCpGs and trans eCpGs. Additionally, we assessed gene ontology among transcripts associated with trans eCpG methylation.

Gene body eCpG analysis. We calculated the number of eCpGs that were negatively correlated with their cognate genes in the following categories: gene body (TSS+/-2,500bp to TES for positive/negative strand genes), intronic, exonic, in first exon, in last exon (as determined by the average exon locations; see supplemental methods).

We next address two hypotheses that aim to explain the presence of negatively and positively correlated eCpGs within gene bodies. The first hypothesis is that negatively correlated gene body eCpGs are the result of intragenic gene regulators (e.g., promoters and enhancers). The second hypothesis posits that positively correlated gene body eCpGs result from the regulation of overlapping genes.

To test the hypothesis that negatively correlated genes are the result of intragenic regulatory elements, we looked for enrichment of negatively correlated vs. positively correlated gene body eCpGs among ChromHMM annotated promoters (states 1-3) or enhancers (states 4-8).

To test the hypothesis that positively correlated gene body eCpGs reside in the promoters of overlapping genes, we first identified eCpGs within our results that were associated with expression transcripts that annotated to overlapping genes. Read-through genes and low-confidence (i.e., LOC, orf, KIAA, FLJ) annotations were excluded. We compared the refseq and ensembl annotations separately. From these eCpGs we were able to compare the numbers of negative and positive eCpG-gene body correlations.

Results

Summary of cohorts and data. We analyzed genome-wide DNA methylomic and transcriptomic data from two cohorts. In the Grady Trauma Project (GTP), whole blood samples were collected from 333 participants (76% female) aged 18 – 78 years (GEO accession numbers GSE72680, GSE58137). In the Multi-Ethnic Study of Atherosclerosis (MESA), relevant data were available for purified monocytes from 1,202 participants (51% female) aged 55 – 94 years (GEO accession number GSE56047, Table 4-1).

For both GTP and MESA, methylation data for >480K individual CpGs were generated from the Infinium HumanMethylation450 BeadChip (Illumina, San Diego, CA), and RNA transcript levels for >25,000 annotated genes were quantified via Illumina HumanHT-12 v3.0 and v4.0 Expression BeadChip (see “Materials and Methods” for details).

Although both studies derive data from blood cells, GTP derives data from whole blood samples, while MESA derives data from purified monocytes (a small component of whole blood cells; see Materials and Methods). As such, we analyze both studies in parallel and make comparisons between the two, but they are not meant to be biological replicates.

General landscape of DNA methylomic profile. We identified 1,687 and 16,327 eCpGs in GTP and MESA respectively (GTP: -53<T<70, 9.7e-197<p<1e-11; MESA: -70<T<54, 1e-321<p<1e-11). These eCpGs associate with 533 and 3,269 transcripts, making a total of 2,466 and 34,518 unique eCpG-transcript pairs for GTP and MESA, respectively (Table 4-2). The discrepancy in the number of findings between GTP and MESA is likely due to power differences; with n=333 for GTP and n=1,202 for MESA and an α-level of 1e-11, the studies have 80% power to detect associations where the eCpG explains as little as 16% (GTP) or 4.7% (MESA) of variation in expression. Another factor that may contribute to the discrepancy is that monocytes have a slightly larger dynamic methylation range than the predominant cell type in whole blood [33]. The average number of eCpGs per transcript was 4.6 and 11 for GTP and MESA, respectively. The median number of eCpGs per transcript was two for both GTP and MESA.

Correlations between methylation and expression were predominantly negative in both GTP (70%, n=2,466) and MESA (53%, n=34,518; Figure 4-2, 4-3A; Table 4-2). For both GTP and MESA, there are more negatively than positively correlated eCpGs among both cis and distal eCpG-transcript pairs. However, while GTP trans eCpGs are enriched for negative correlations (OR=1.6, P=3.9e-7), MESA trans eCpGs are enriched for positive eCpG-transcript pair correlations (OR=1.6, P<2.2e-16; Table 4-2).

Sets of assayed CpGs within GTP and MESA displayed the expected bimodal distribution of average methylation values, suggesting that most CpGs were either fully methylated or unmethylated. In contrast, eCpGs were more likely to be intermediately methylated, with average β-values between 0.2 and 0.8 (OR=3.6 (MESA), 3.04 (GTP), Fisher's exact P<2.2e-16 for both MESA and GTP; Figure S4-1). This relationship may simply reflect increased power due to increased variability among intermediately methylated CpGs. Consistent with this, variability in β-values is greater in eCpGs than non-eCpGs in both GTP and MESA (Figure S4-2).

Distribution of eCpGs relative to the 450K array. When on the same chromosome (cis and distal), eCpGs were located in the associated gene or within 2,500bp of its TSS 49% (n=1,508) and 41% (n=10,706) of the time, for GTP and MESA, respectively. However, we find that the relative proportion of eCpGs ([number eCpGs per bin/total number eCpGs] / [number CpGs per bin/total number CpGs]) increases with proximity to the associated transcript, but drops dramatically very near and in the transcript (Figure 4-3B). Accordingly, the proportion of eCpGs distal to their associated (or cognate) gene exceeds the proportion of CpGs on the array that are distal to the closest transcript (for CpGs and transcripts passing QC in each study; Figure 4-3C). There also appears to be a predominance of eCpGs located upstream of their associated gene (Figure 4-2, third column); however, this imbalance reflects the composition of the Human-Methylation450 array (Figure S4-3).

Distribution of eCpGs relative to associated genes. In GTP and MESA, distal and trans eCpGs constitute 53% (n=2,466) and 79% (n=34,518) of eCpGs, respectively (Figure 4-4, Table 4-2, S4-1), indicating that eCpGs are not primarily near associated genes. Figure 4-1 defines the possible eCpG-transcript pair scenarios, relative to the gene annotated to the transcript and other nearby genes, described further in Materials and Methods. In short, we consider canonical eCpG-transcript pairs to be those in which the eCpG is within the gene or within 2,500 bp of the gene's TSS, or the associated gene is the closest gene to the eCpG. Among cis eCpGs, nearly 35% do not conform to a canonical methylation-expression role where the eCpG associates with the nearest gene (GTP n=1,167; MESA n=7,246; Figure 4-4). Canonical eCpG-transcript pairs are captured in the remaining 65% of cis eCpGs (GTP n=1,167; MESA n=7,246; 21 to 48% of all eCpGs; GTP n=2,466; MESA n=34,518; see Figure 4-4).

Corroboration of eCpG results

Between study comparison. To corroborate the eCpGs identified here, we compared eCpG-transcript pairs across studies. Among eCpG-transcript pairs significant in GTP, 44% (n=1,260) of cis pairs (53% of promoter eCpG-transcript pairs, n=383), and 30% of distal (n=341) and 27% trans (n=958) pairs are significant in MESA. Randomly permuting the transcript IDs among the significant eCpG-transcript pairs from both studies and repeating the calculation 10,000 times yielded no higher than 3% of GTP distal- and trans- pairs occurring in MESA distal- and trans- pairs.

Within study comparison. To corroborate our eCpGs within each study, we examined associated gene congruence among neighboring eCpGs. Among eCpGs having a neighbor within 500bp, 97% (GTP) and 90% (MESA) have a neighbor significantly associated with at least one of the same cognate genes, 86% (GTP) and 89% (MESA) have at least one neighbor that is consistent with regard to direction of correlation and 82% (GTP) and 87% (MESA) have completely congruent neighbors (GTP n=738, MESA n=6,290; Figure 4-5). Trans eCpGs have a slightly lower proportion of neighbors significantly associated with the same cognate gene (GTP=91%, MESA=88%), but 85% of GTP and 88% of MESA eCpGs have all neighbors of congruent direction (GTP n=109, MESA n= 3,074). The proportion of proximal CpG neighbors with matching associated gene and sign predictably declines with increasing window size (Figure 4-5).

Functional analysis of eCpGs

Functional trends among all eCpGs. Next we used publicly available data to assess functional trends among eCpGs [28, 29, 34]. As part of the ENCODE project [28] Ernst, et al. (2011) used a hidden Markov model to partition the genome into functional domains based on ChIP-seq data for histone modifications, RNA polymerase occupancy, and other chromatin features. We used the resulting data set, called ChromHMM, along with CpG island, long intergenic non-coding RNA (lincRNA), transcription factor binding site (TFBS) and small nucleolar and microRNA (sno/microRNA) genomic intervals to evaluate the chromatin structure surrounding eCpGs [28, 29, 34]. When considering all CpGs tested in MESA, genome-wide significant eCpGs are depleted among CpG islands (CGI; OR=0.60, P=7.5e-170) and promoters (ChromHMM states 1-3; OR=0.55, P=1.4e-168), but enriched among the more variable CpG shore (1,500bp out from CGI; OR=1.2, P=2.1e-22) and shelf (1,500bp out from CG shores; OR=1.2, P=2.6e-12) regions (Figure 4-6A). We also find that eCpGs are enriched among transcription factor binding sites (TFBS) and highly enriched among annotated enhancer regions (ChromHMM states 4-7; Figure 4-6A; OR>1.9 and P<2.2e-16). GTP shows a similar enrichment for enhancer regions (Figure S4-4, top row). This result is consistent with other studies that have found a significant enrichment of eCpGs among enhancers [18, 23].

Functional trends among Trans eCpGs. When assessing the enrichment of chromatin states among the various categories of significant eCpGs, we find that trans eCpGs are enriched among both strong (ChromHMM states 4 and 5; OR=1.8, P=2.2e-72) and weak (ChromHMM states 6 and 7; OR=1.4, P=5.4e-18) enhancer annotations (Figure 4-6B). Like the overall pattern, we see that trans eCpGs are depleted among CGI (OR=0.83, P=3.6e-10) and promoters (OR=0.45, P=5.5e-168), and enriched among TFBS (OR=1.3, P=2.8e-21). Interestingly, we also observe trans eCpGs to be enriched among regions of the genome that are annotated as sno and microRNAs (OR=2.4, P=9.8e-3; Figure 4-6B).

Functional trends among Cis and distal eCpGs. Unlike trans eCpGs, we see that enhancers are primarily depleted among the various cis and distal eCpG categories described in Materials and methods and Figure 4-1. Additionally, insulators (ChromHMM state 8) are enriched among cis and distal eCpGs (1.4<OR<8.2, P<0.03). Promoter (1.1<OR<6.6, P<0.04) and CpG islands 1.3<OR<1.8, P<0.03, shores (1.6<OR<3.2, P<4.8e-07) and shelves (1.4<OR<1.6, P<0.01) are more often enriched among cis eCpG categories. We also see a strong enrichment of cis (1.7<OR<2.5, P<0.05) and distal (OR=1.4, P=5e-4) eCpGs among regions of the genome annotated as lincRNAs. We note a depletion of enhancers in the cis categories in which lincRNA eCpGs are enriched (OR=0.5, P=5.9e-05; Figure 4-6B).

eCpG-transcript distances suggest possible DNA looping. To interrogate the nature of the connection between eCpGs and their associated transcripts, we assessed the distribution of distances between them. We compared the distribution of eCpG-transcript distances to the distribution of DNA looping interaction distances, as captured by HiC. We considered eCpG-transcript frequencies that decrease as a function of distance to be consistent with the DNA looping frequencies seen in HiC data [30]. We find a strong negative correlation between the number of eCpG-transcript pairs and the distance separating them (Figure 4-7, green). A similar decay is seen in the distribution of DNA looping interactions captured by HiC (Figure 4-7, blue).

Gene Ontology analysis. We used GO to assess molecular function terms among all eCpGs, cis and distal eCpGs and trans eCpGs, as well as among transcripts associated with trans eCpG methylation. We found that eCpGs are enriched for nucleotide binding molecular functions, like sequence specific DNA binding (OR=2.6, P=1.2e-04) and transcription factor binding (OR=4.3, P=2.6e-04). DNA-binding and transcription factor molecular functions are also enriched in cis/distal and trans eCpGs (1.5<OR<4.2, P<1.8e-04). Finally, transcripts that associate with trans eCpG methylation were enriched for chromatin readers, writers (1.7<OR<3.8, P<6.3e-04) and transcription co-activator genes (Ligand-dependent nuclear receptor transcription coactivator activity OR=2.9, P=1.2e-03; Tables S4-2-S4-5). All p-values listed above correspond with a false discovery rate (FDR)<0.05

Analysis of gene body eCpGs. It has been frequently reported that DNA methylation is negatively correlated with gene expression in promoters, but positively correlated with gene expression within gene bodies [2, 18, 21]. Here, we observe that DNA methylation is negatively correlated with transcript expression the majority of the time, in any location (Figure 4-8). Among significant eCpGs in MESA, negative correlations are enriched among gene body eCpGs (OR=1.5, P=2.6e-16). Among significant eCpG-transcript associations where the CpG was located within the gene body of its associated transcript, 1) the correlation was negative 71% (n=356) and 62% (n=1,919) of the time, for GTP and MESA, respectively, 2) the direction of correlation was consistent across multiple eCpGs within a single transcript 85% (n=87; GTP) and 72% (n=601; MESA) of the time, and 3) among transcripts with consistent associations across multiple eCpGs, the correlations were negative 81% (n=74; GTP) and 76% (n=434; MESA) of the time. Among CpGs within the first and last exon of their associated transcript, we note that although still primarily negative, fewer eCpGs are negatively correlated with transcript expression in the last exon (59% in MESA, 73% in GTP), in comparison to the first exon (77% in MESA, 87% in GTP) (Figure 4-8).

Functional trends among gene body eCpGs that negatively correlate with expression. One hypothesis that attempts to account for an excess of negative correlations among gene body eCpGs posits that these eCpGs are found in intragenic regulatory elements like promoters and enhancers located within the gene they control [35, 36]. We observe a slight enrichment of annotated promoters among negatively correlated gene body eCpGs (OR=1.8, P=0.002). An even stronger enrichment of negatively correlated gene body eCpGs among annotated enhancer regions (OR=2.2, P<2.2e-16) suggests that transcriptional regulators within gene bodies may be important to gene regulation.

Competing promoters among gene body eCpGs that positively correlate with expression. A separate hypothesis states that the presence of positively correlated gene body eCpGs may result from the regulation of an overlapping gene [37, 38]. To test this hypothesis, we found eCpGs that are within the promoter of one gene and the gene body of another. We found five eCpGs that fit these criteria. In three of five cases, the eCpG correlated negatively with the transcript when the eCpG was near the promoter and positively with the transcript when the eCpG was in the gene body. In the remaining cases, the eCpG correlated negatively with both transcripts (Table S4-6).

Discussion

EWAS often identify CpGs that lie outside of defined genomic regions like promoters, which are typically considered the canonical target for epigenetic gene repression [2, 16, 17]. Inferring a functional consequence for these CpGs is difficult because our understanding of the role of methylation in regulation of gene expression and disease is incomplete. We find that the majority of eCpGs do not conform to canonical methylation-expression roles. Our results highlight a shortcoming of current CpG functional annotation, as these non-canonical methylation-expression relationships would be incorrectly assigned to the nearest gene in EWAS interpretation.

We find that many eCpG-transcript pairs are consistent between studies and that neighboring eCpGs within studies tend to correlate with the same gene. Although it is encouraging to find matching pairs between studies, it is unsurprising that there is not complete overlap given differences in both power and cell type and ethnic background across studies. GTP is a relatively small study, whose data were derived from whole blood in an African American cohort. MESA, a much larger study from a cohort of mixed ethnicity, derived data from monocytes, which only account for a small proportion of whole blood cells, on average. As such, MESA and GTP are not intended to be replicates but a comparison across whole blood and monocytes. In a study of cis CpG-transcript associations, Liu et al. (2013) found that few observed expression-associated methylation sites were specific to any ethnic category, so it is unlikely that differences between eCpGs found in GTP and MESA are driven by ethnic composition. Our results suggest that our eCpGs represent robust associations that are consistent between neighboring CpGs and across datasets.

Among transcripts passing QC (GTP: 13,933, MESA: 19,445), only 3.8% of GTP transcripts and 17% of MESA transcripts significantly correlated with CpG methylation. Because the two studies are powered to detect associations explaining >16% or >4.7% of variance in expression, respectively, eCpG-transcript associations with subtler correlations would not have been detected. It is possible that in many cases either the transcripts or the CpGs passing QC were not variable enough in the tissues studied to detect associations, or that some of the genes are not epigenetically regulated in blood. This hypothesis is supported by the observation that in MESA, which is powered to detect subtler associations than GTP, the average variance in methylation β-values for identified eCpGs was lower (3.6e-03) than for eCpGs identified in GTP (6.4e-03), while variation in non-eCpGs was similar across both datasets (Figure S4-2). Finally, the variance in some genes could be due to factors other than CpG methylation, for instance, regulation by other genes or higher-level chromatin mark (i.e. histone modifications).

Our enrichment and gene ontology results make the case for a complex network of epigenetic control. In addition to the more canonical promoter eCpGs that associate with proximal gene expression, we also see that eCpGs associate with gene expression distally, through enhancers, insulators and long intergenic non-coding RNAs (lincRNAs). Importantly, we find that enhancer elements, micro and small nucleolar RNAs are prominent among eCpGs that correlate with the expression of genes on different chromosomes (trans). The GO analysis suggests that for each gene, we have likely constructed a regulatory profile that encompasses the indirect, trans effects (which could include regulatory networks) as well as direct, cis effects (including cis and distal DNA methylation). Because we find many eCpGs, genome-wide, that associate with transcription factor genes and chromatin modifiers, our results may include scenarios in which gene expression influences DNA methylation patterns, as well as vice-versa [39]. Although these findings represent associations and do not provide information on causality; they could prove useful in annotating EWAS results for CpGs with potential roles in regulatory networks.

Overall, our results indicate that CpG methylation interacts with gene expression primarily through enhancer CpGs, rather than promoter CpGs. Enhancers, as distal regulatory elements, are methylation sensitive transcription factor binding sites that promote tissue-specific gene expression [2, 3]. Other studies have also noted an enrichment of enhancer regions among eCpGs [18, 23]. One proposed model of gene regulation suggests that promoter methylation is relatively static, having either a restrictive (hypermethylated) state, or permissive (hypomethylated) state at which dynamic enhancer methylation modulates gene expression levels [18]. In this scenario, promoter eCpGs are far less likely than enhancer eCpGs to be identified due to their low variability [23]. Our results support the important role of enhancer CpG methylation in epigenetic gene regulation, but expand on this model to suggest that enhancer methylation can correlate with gene expression changes on other chromosomes.

We also find that insulator eCpG methylation plays a prominent role in cis and distal gene expression. Insulators are thought to promote gene expression by bringing enhancers and promoters into close proximity through the binding of the CCCTC-binding factor (CTCF), which can dimerize to form stable chromatin loops [30, 40]. The binding affinity of CTCF to insulator sequences is influenced by DNA methylation [41]. Here we see that insulators are enriched among cis and distal eCpGs. We also see that the frequency of eCpG-transcript interactions decreases with distance, as seen in the DNA looping interaction frequency captured by HiC [30]. Overall, our results support the role of insulators in regulation of gene expression, potentially through the formation of functional DNA loops involving enhancer and insulator elements.

MicroRNAs regulate more than 50% of mRNAs [42] and are in turn regulated by DNA methylation [43, 44]. We see a strong enrichment of trans eCpGs among micro/snoRNAs, so it is intriguing to speculate that trans eCpG-transcript associations are due, at least in part, to post-transcriptional regulation by microRNAs. We also see that cis and distal eCpGs are enriched among lincRNAs. Evidence suggests that lincRNAs play an important role in gene expression, particularly as eRNAs (enhancer RNAs), which are RNAs transcribed from enhancer sequences and may act as scaffolding for DNA looping or co-activator recruitment to a gene promoter [40]. Interestingly, the enhancers that give rise to eRNAs are distinct from enhancers that act as transcription factor binding sequences [45]. In our results, we also see a depletion of enhancers in the cis categories in which lincRNA eCpGs are enriched. From our results, we propose that DNA methylation may be a key player in cis, distal and trans transcriptional control through the action of non-coding RNAs.

Our study finds that most eCpG-transcript correlations are negative, even among gene bodies. Our findings are in line with other studies that report the predominance of negative correlations [19, 20, 22, 23]. The primary difference between studies that find mostly negative methylation-expression correlations and those that find negative correlations in promoters and positive correlations in gene bodies is study design. Most studies finding positive gene body correlations were considering the correlation of expression and methylation across all genes in a single genome [18, 46, 47]. In contrast, the majority of studies finding negative correlations in gene bodies were considering correlation of expression and methylation across individuals, separately for each CpG [21, 23]. A within-genome comparison observing that more highly expressed genes tend to show hypermethylation within gene bodies is simply a comparison of different genes and does not speak to the effect of changes in DNA methylation at any particular gene. In general, studies that assess DNA methylation in gene bodies across individuals find that, most of the time, increases in DNA methylation are associated with decreases in gene expression [19, 20, 22, 23].

We provide evidence here that negatively correlated gene-body eCpGs are often the result of intragenic regulatory elements (e.g. promoters and enhancers). We also provide support for the hypothesis that positive correlations between CpG methylation and gene expression are the result of overlapping genes/variants [37, 38]. Neither of these hypotheses fully explain the occurrence of either positive or negative eCpG correlations within gene bodies. Rather, they suggest that there is no all-encompassing biological truth to these associations.

Conclusions

We have characterized the genome-wide DNA methylomic profile for gene expression in human blood cells. Many of our results are reproducible between whole blood and monocytes and are spatially correlated within studies. Unlike similar studies, we found that most eCpGs were very distal and trans to their associated genes. These results highlight the shortcomings of proximity based CpG annotations, as even cis eCpG-transcript associations often do not involve the closest downstream TSS. In fact, the majority of associations were distal or trans, representing a serious gap in functional annotation for epigenome-wide association studies.

Like others, we find an overabundance of enhancer eCpGs, highlighting the importance of enhancers, possibly over promoters, in gene expression variation [18, 23]. We also note enrichments of insulators and non-coding RNAs, like microRNAs and lincRNAs among eCpGs. Our results point to DNA methylation as a possible link between gene expression and higher-order chromatin organization, as well as another layer in post-transcriptional regulation.

Like studies of similar design, we find an abundance of negative CpG-transcript associations [19, 20, 22, 23], which conflicts with earlier reports that gene body methylation positively correlates with gene expression [18, 36, 46, 48, 49]. We find some support for the hypothesis that negatively-correlated gene-body eCpGs are in annotated promoters and enhancers [36], which suggests an important role for alternate gene-body promoters and intragenic enhancers in gene expression. However, we do not find support for the presence of negative gene-body methylation associations as a result of overlapping gene expression.

Finally, our gene ontology results, like our enrichment results, portray a complex, multi-dimensional picture of epigenetic interactions in the genome. eCpGs are enriched in molecular functions like transcription factor binding and sequence specific DNA binding. Among transcripts that associate with trans eCpG methylation, we find an enrichment of chromatin readers, writers and transcription co-activator genes.

Our findings suggest that limiting our interpretation of EWAS results to the nearest gene might be short-sighted, as DNA methylation may have many indirect interactions (e.g. modulating the expression of a transcription factor) that influence gene expression or vice-versa. Overall, our results broaden our understanding of the ways that CpG methylation interacts with gene expression, genome-wide, and provide data that may be useful for mining meaningful biological insights from EWAS.

Tables

Table 4-1. Cohorts and Data for GTP and MESA

	GTP	MESA
Participants	333	1,202
Tissue	Whole blood	Monocytes
Original study phenotype	Post traumatic stress disorder	Atherosclerosis
Methylation technology	Infinium HumanMethylation450 BeadChip	
Expression technology	Illumina HumanHT-12 Expression BeadChip	
Methylation probes included	472,199	422,016
Expression probes included	13,933	19,445

*$p \leq 10^{-11}$

Table 4-2. eCpG results for GTP and MESA

Study	GTP			MESA		
Number of eCpGs	1,692			16,356		
Number of transcripts	537			3,277		
eCpG-transcript pairs	2,466			34,518		
Transcript pair status	Cis	Distal	Trans	Cis	Distal	Trans
Total pairs	1,167	341	958	7,246	3,460	23,812
Positively correlated	389	114	228	2,560	1,578	11,985
Negatively correlated	778	227	730	4,686	1,882	11,827

*p $\leq 10^{-11}$

Figures

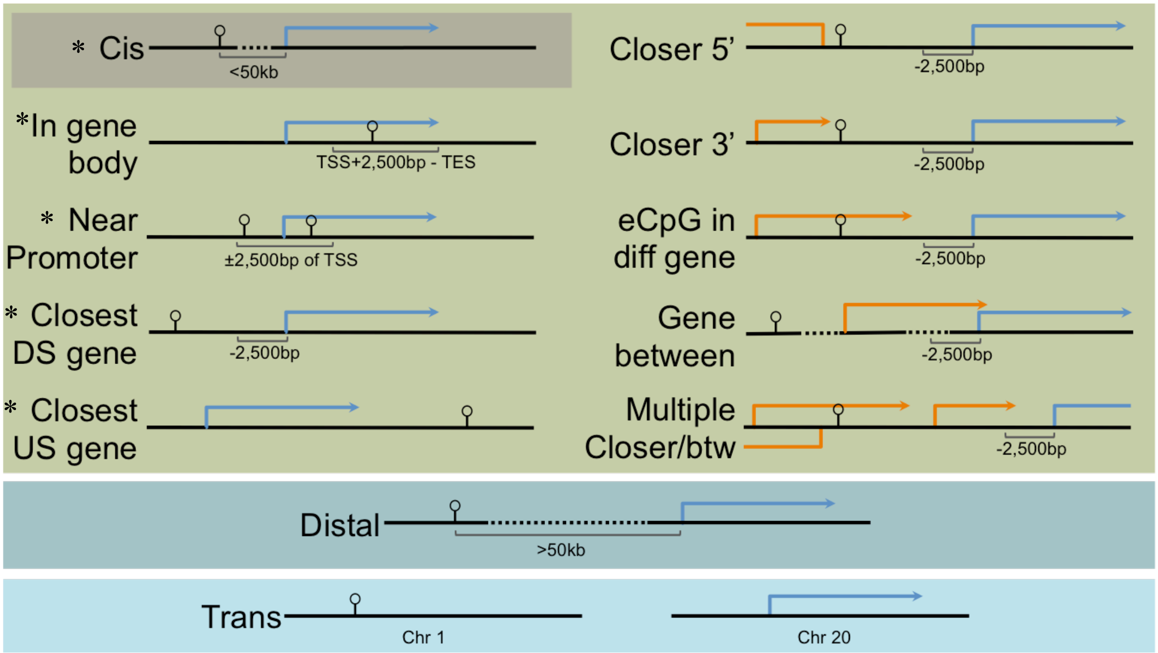


Figure 4-1. Graphical examples of each functional category. Examples are shown for positive strand eCpG (stick with open circle) and transcript (blue arrow) pair associations. Blue arrows represent the gene transcription area (TSS-TES) that was annotated to the expression probe in the eCpG-transcript pair by overlap with a refseq or ensemble exon. Orange arrows represent examples of other annotated genes that are near the eCpG-transcript pair. DS is downstream, US is upstream, TSS is transcription start site, and TES is transcription end site. * indicates canonical methylation-expression roles.

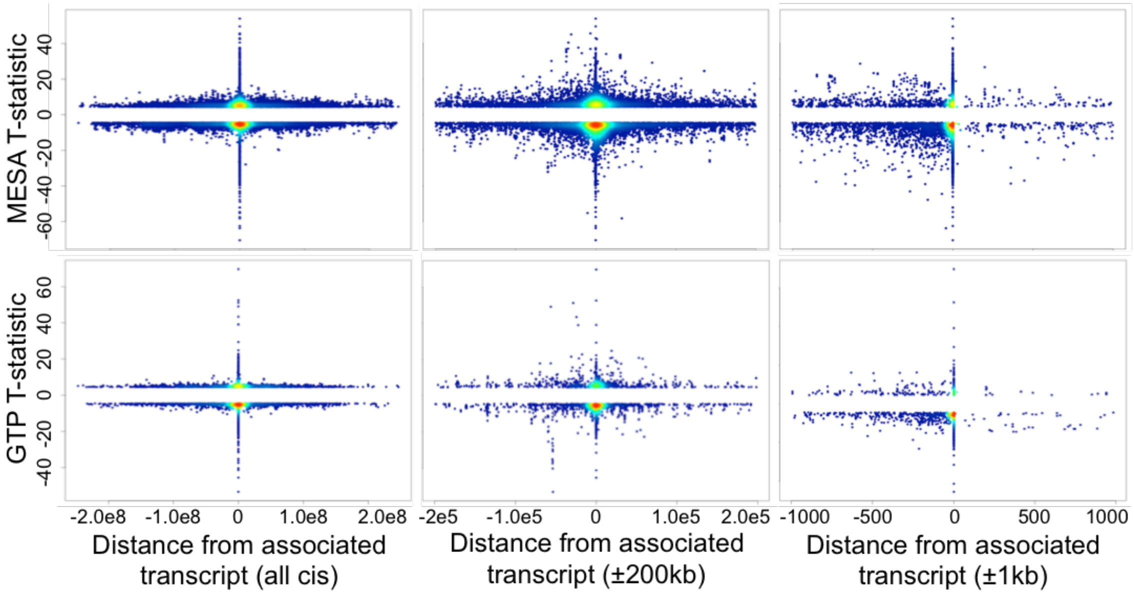


Figure 4-2. Scatter plot of T-statistic vs. distance from associated transcript, among suggestively significant eCpGs (P<1e-5). The top row is from MESA and the bottom from GTP. The leftmost column is for all cis and distal eCpGs. The middle and right columns contain only eCpGs within 200kb and 1kb from their cognate transcript, respectively.

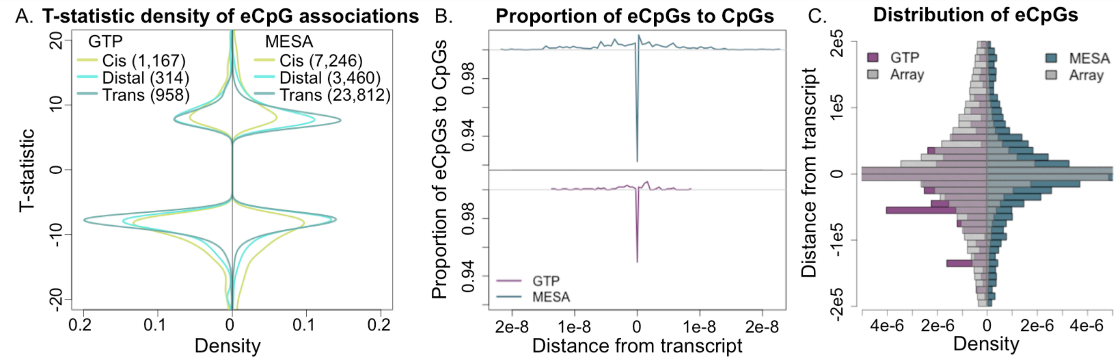


Figure 4-3. Distribution of genome-wide significant eCpGs (P<1e-11). More negative than positive associations are seen in both studies (A). The proportion of CpGs that are eCpGs rises near genes, but drops very near and in their associated genes (B). eCpGs are found distal to their associated genes (C).

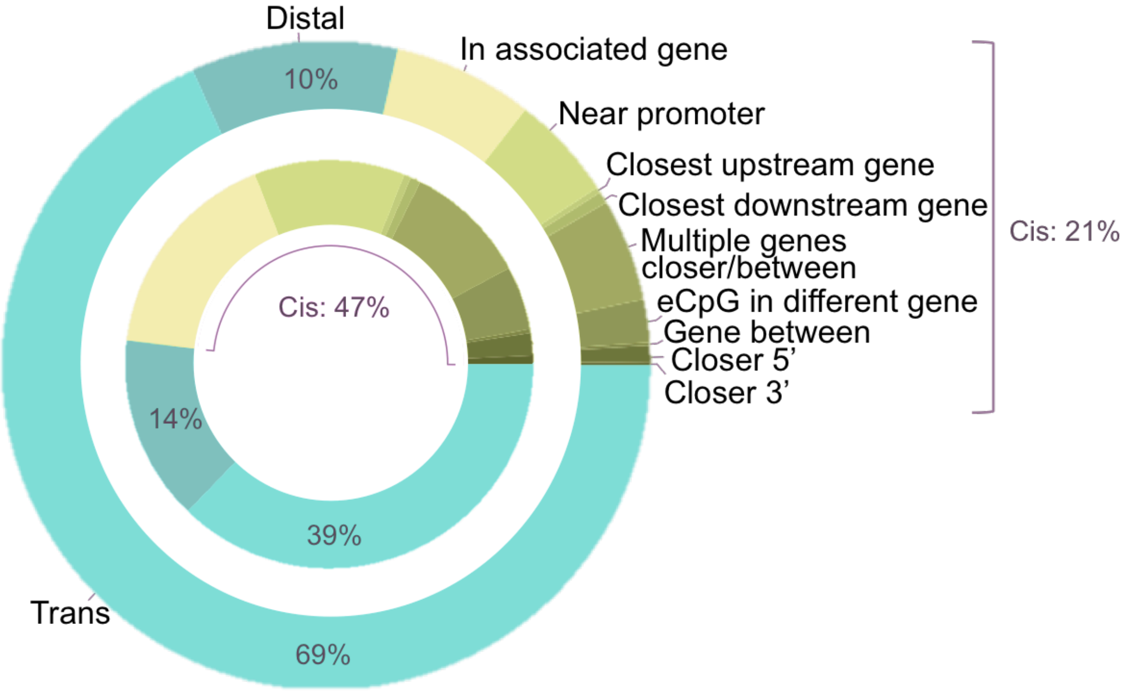


Figure 4-4. Genome-wide significant (P<1e-11) eCpG-transcript relationship proportions in GTP (inner; n=2,466) and MESA (outer; n=34,518). The green sections represent eCpGs that are <50kb from their associated transcript (cis); yellow represents eCpGs that fall within the gene body of their associated transcript; dark blue represents eCpGs that were <50kb, but on the same chromosome (distal) as the associated transcript; and light blue represents eCpGs that were on a different chromosome from the associated transcript (trans). Definitions of each category are given in Figure 4-1 and Materials and methods section.

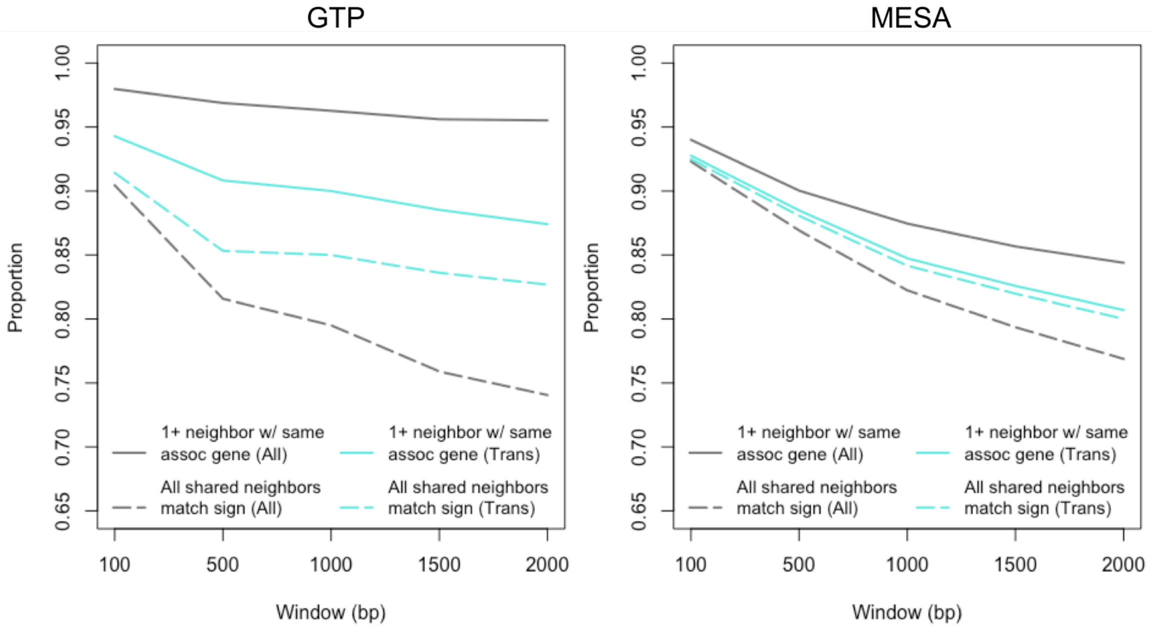
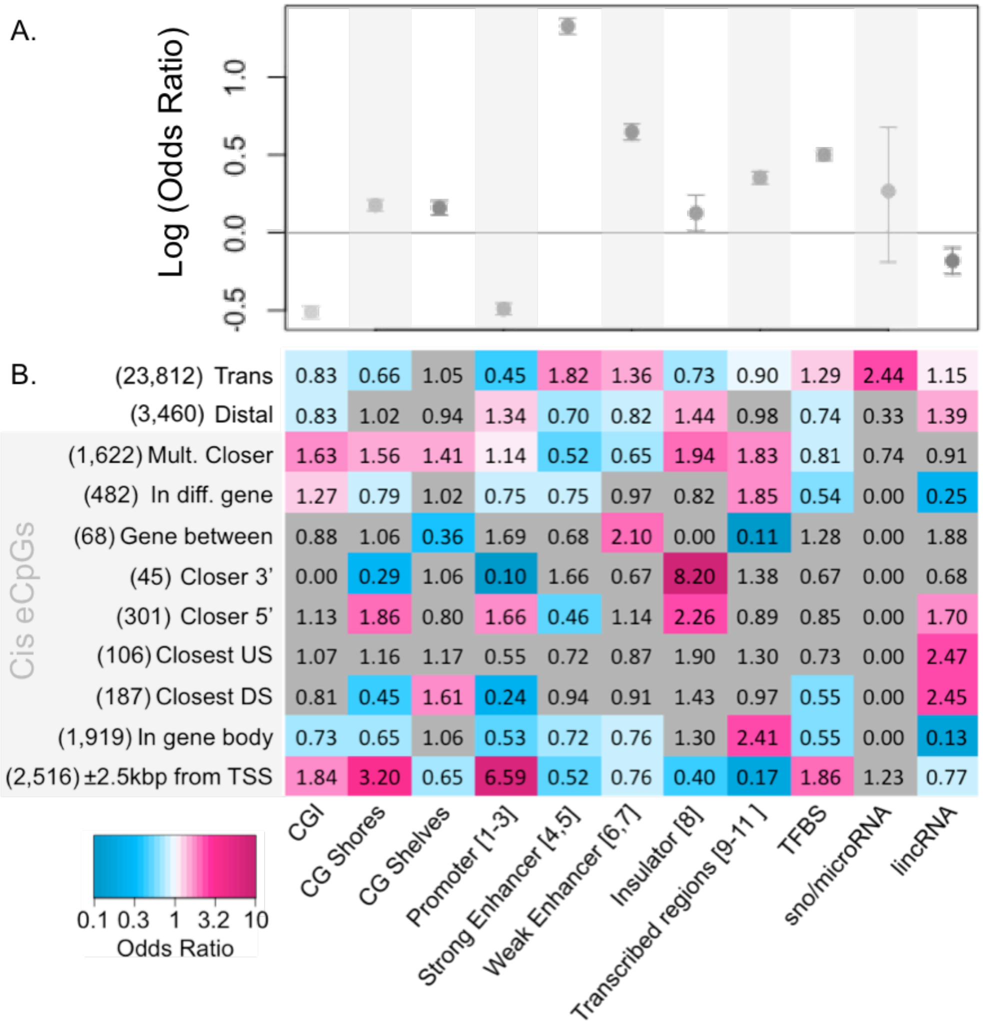


Figure 4-5. Shared gene associations among neighboring eCpGs. Proportion of proximal eCpGs (neighbors) in GTP and MESA with the same associated gene or same associated gene and direction of association as the query CpG. Neighbors were located within the specified window size on either side of the query CpG. Associated gene overlap among proximal eCpGs appears to be a function of distance. The majority of neighboring eCpGs sharing an associated gene, associate with the gene in the same direction.



	CGI	CG Shores	CG Shelves	Promoter [1-3]	Strong Enhancer [4,5]	Weak Enhancer [6,7]	Insulator [8]	Transcribed regions [9-11]	TFBS	sno/microRNA	lincRNA
(23,812) Trans	0.83	0.66	1.05	0.45	1.82	1.36	0.73	0.90	1.29	2.44	1.15
(3,460) Distal	0.83	1.02	0.94	1.34	0.70	0.82	1.44	0.98	0.74	0.33	1.39
(1,622) Mult. Closer	1.63	1.56	1.41	1.14	0.52	0.65	1.94	1.83	0.81	0.74	0.91
(482) In diff. gene	1.27	0.79	1.02	0.75	0.75	0.97	0.82	1.85	0.54	0.00	0.25
(68) Gene between	0.88	1.06	0.36	1.69	0.68	2.10	0.00	0.11	1.28	0.00	1.88
(45) Closer 3'	0.00	0.29	1.06	0.10	1.66	0.67	8.20	1.38	0.67	0.00	0.68
(301) Closer 5'	1.13	1.86	0.80	1.66	0.46	1.14	2.26	0.89	0.85	0.00	1.70
(106) Closest US	1.07	1.16	1.17	0.55	0.72	0.87	1.90	1.30	0.73	0.00	2.47
(187) Closest DS	0.81	0.45	1.61	0.24	0.94	0.91	1.43	0.97	0.55	0.00	2.45
(1,919) In gene body	0.73	0.65	1.06	0.53	0.72	0.76	1.30	2.41	0.55	0.00	0.13
(2,516) ±2.5kbp from TSS	1.84	3.20	0.65	6.59	0.52	0.76	0.40	0.17	1.86	1.23	0.77

Figure 4-6. Enrichment of chromatin features among eCpGs. A) Enrichment of eCpGs (odds ratios and 95% confidence intervals) for the listed chromatin features, among all CpGs tested in MESA (N=422,016). **B)** Enrichment of eCpGs for the listed chromatin features, among genome-wide significant eCpGs in MESA (N=34,518). Shaded categories are cis. Blue indicates significant depletion and red, significant enrichment (P<0.05). Bracketed numbers in the chromatin features indicate the ChromHMM state. Numbers in parentheses indicate the number of eCpGs in the category. **Definitions: Left: Given in Figure 4-1 and Materials and Methods Bottom.** "CGI" are CpG islands. "TFBS" is transcription factor binding site.

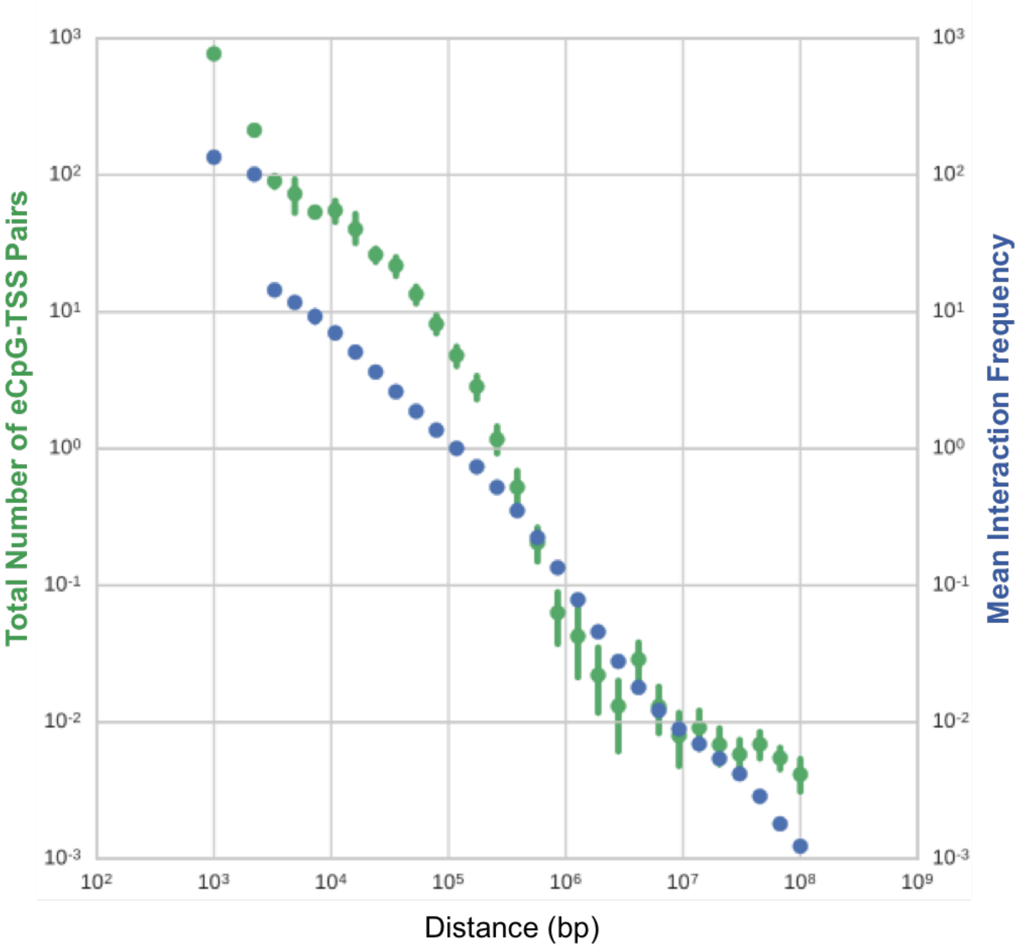


Figure 4-7. eCpG-transcript distance vs. HiC interaction frequency. Distribution of interactions over distance for MESA cis and distal eCpG-transcript pairs (green) and Hi-C interaction data from a lymphoblastoid cell line (GM12878; blue). Mean interaction frequency between each 1 kb bin (blue dots) were calculated for unnormalized Hi-C interaction counts on Chromosome 1. Number of eCpG-transcript pairs (green dots) per 1 kb were calculated by rounded distance between eCpG and TSS for cis and distal eCpG-transcript pairs. Bars represent 0.95 confidence intervals for each bin. The decay curve of eCpG-transcript pair distances within 10^6 bp is consistent with the chromatin looping interaction curve seen in HiC.

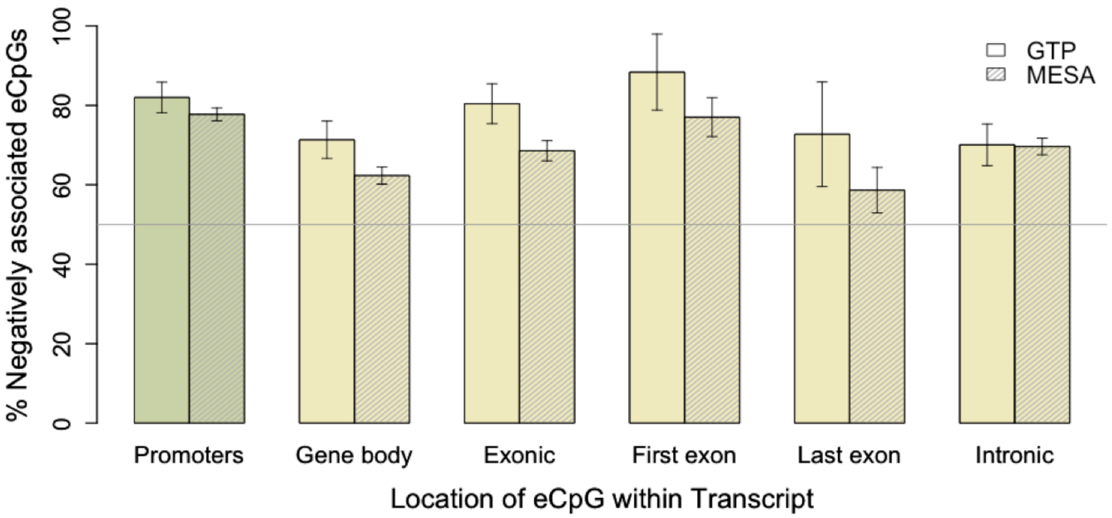


Figure 4-8. Negative eCpG-transcript correlations in GTP and MESA. The fraction of negative eCpG-transcript associations is greater than 50% in promoters and gene bodies. More negative associations are found in the first exon than the last.

Supplemental tables

Table S4-1. Breakdown of eCpG-transcript status

MESA		
eCpG-transcript status	Counts	Percent
Cis	7,246	20.99
In gene body	1,919	26.48
± 2,500bp of TSS	2,516	34.72
Closest upstream gene	106	1.46
Closest downstream gene	187	2.58
Multiple genes closer/between	1,622	22.38
eCpG in different gene	482	6.65
Gene between	68	0.94
Closer 5’	301	4.15
Closer 3’	45	0.62
Distal	3,460	10.02
Trans	23,812	68.98
Total	34,518	

GTP		
eCpG-transcript status	Counts	Percent
Cis	1,167	47.32
In gene body	356	30.51
± 2,500bp of TSS	383	32.82
Closest upstream gene	17	1.46
Closest downstream gene	18	1.54
Multiple genes closer/between	243	20.82
eCpG in different gene	81	6.94
Gene between	7	0.60
Closer 5’	49	4.20
Closer 3’	13	1.11
Distal	341	13.83
Trans	958	38.85
Total	2,466	

Table S4-2. Gene Ontology enrichment among all eCpGs

Term (MF)	Odds Ratio	P-value	FDR
Sequence-specific DNA binding	2.64	1.16E-04	3.23E-02
Ribonucleoside binding	1.45	1.17E-04	3.23E-02
Purine nucleoside binding	1.45	1.21E-04	3.23E-02
Purine ribonucleotide binding	1.43	1.72E-04	3.46E-02
Binding	1.51	1.94E-04	3.46E-02
Transcription factor binding	4.31	2.56E-04	3.92E-02

Table S4-3. Gene Ontology enrichment among cis and distal eCpGs

Term (MF)	Odds Ratio	P-value	FDR
Actin binding	1.66	2.57E-05	1.94E-02
Actin filament binding	2.37	5.15E-05	1.94E-02
Ankyrin binding	10.35	5.26E-05	1.94E-02
HMG box domain binding	16.33	1.41E-04	3.91E-02
Sequence-specific DNA binding	1.48	1.80E-04	3.98E-02

Table S4-4. Gene Ontology enrichment among trans eCpGs

Term (MF)	Odds Ratio	P-value	FDR
Sequence-specific DNA binding	2.99	4.56E-06	2.45E-03
Ribonucleoside binding	1.46	2.12E-05	5.75E-03
Purine nucleoside binding	1.46	2.19E-05	5.75E-03
Purine ribonucleotide binding	1.45	2.68E-05	5.75E-03
Transcription factor binding	4.22	6.37E-05	1.14E-02
Ligase activity	2.88	1.08E-04	1.66E-02
Double-stranded DNA binding	Inf	1.35E-04	1.71E-02
Protein homodimerization activity	1.79	1.43E-04	1.71E-02
Adenyl nucleotide binding	1.43	1.96E-04	2.11E-02
Rab GTPase binding	Inf	2.28E-04	2.22E-02
Protein domain specific binding	3.81	2.79E-04	2.32E-02
ATP binding	1.42	2.81E-04	2.32E-02
Sequence-specific DNA binding transcription factor activity	1.70	3.32E-04	2.55E-02
Binding	1.43	4.08E-04	2.92E-02
Protein kinase binding	2.23	5.58E-04	3.74E-02
RNA polymerase II core promoter proximal region sequence-specific DNA binding	3.58	5.99E-04	3.78E-02

Table S4-5. Gene Ontology enrichment among trans eCpG associated transcripts

Term (MF)	Odds Ratio	P-value	FDR
DNA binding	1.39	1.08E-06	1.17E-04
Single-stranded DNA binding	3.33	3.02E-06	2.94E-04
Hydrolase activity, acting on acid anhydrides	1.50	4.22E-06	3.74E-04
Structural constituent of ribosome	2.52	6.54E-06	5.32E-04
Ubiquitin-protein transferase activity	1.82	1.34E-05	9.50E-04
Chromatin binding	1.71	1.36E-05	9.50E-04
Ligase activity	1.73	6.21E-05	4.03E-03
Ubiquitin binding	3.73	1.43E-04	8.70E-03
ATP-dependent helicase activity	3.86	1.60E-04	9.16E-03
Helicase activity	2.47	2.31E-04	1.25E-02
Histone acetyltransferase activity	3.78	3.92E-04	2.01E-02
RNA binding	1.68	5.66E-04	2.66E-02
Nucleosomal DNA binding	4.32	5.72E-04	2.66E-02
Protein transporter activity	2.48	6.15E-04	2.66E-02
Histone deacetylase binding	2.39	6.27E-04	2.66E-02
Protein C-terminus binding	1.74	1.07E-03	4.35E-02
Ligand-dependent nuclear receptor transcription coactivator activity	2.85	1.17E-03	4.57E-02

Table S4-6. Overlapping gene regulation

First gene	eCpG location	Strand	Corr	Second gene	eCpG location	Strand	Corr
TYMP	Gene body	-	+	*SCO2*	Promoter	-	-
TMIGD3	Gene body	-	+	*ADORA3*	Promoter	-	-
HLA-DPA1	Gene body	-	+	*HLA-DPB1*	Promoter	+	-
KRT10	Gene body	+	-	*TMEM99*	Promoter	+	-
GFM1	Gene body	+	-	*LXN*	Promoter	-	-

Supplemental figures

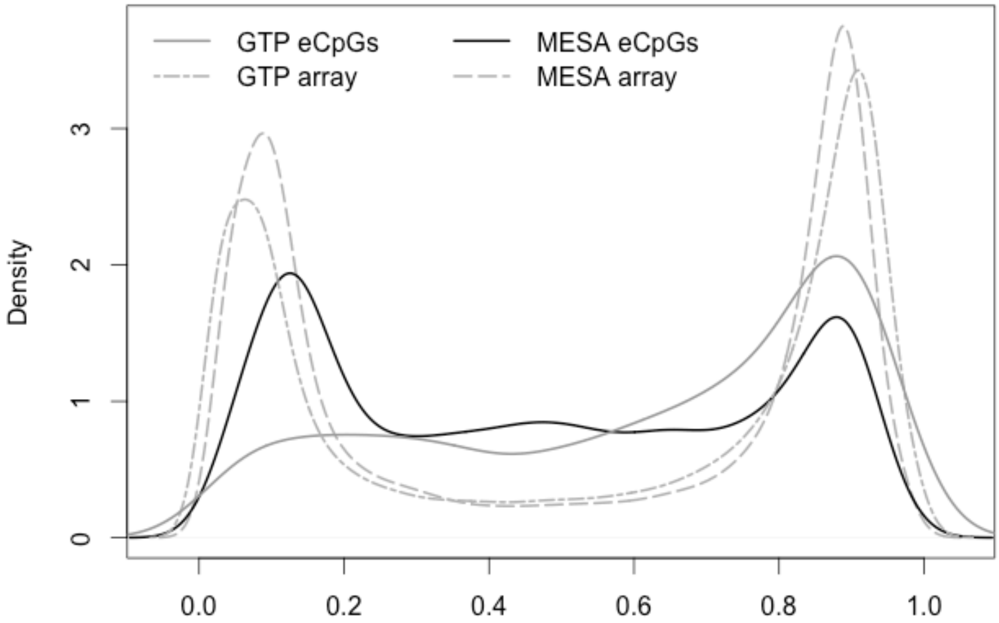
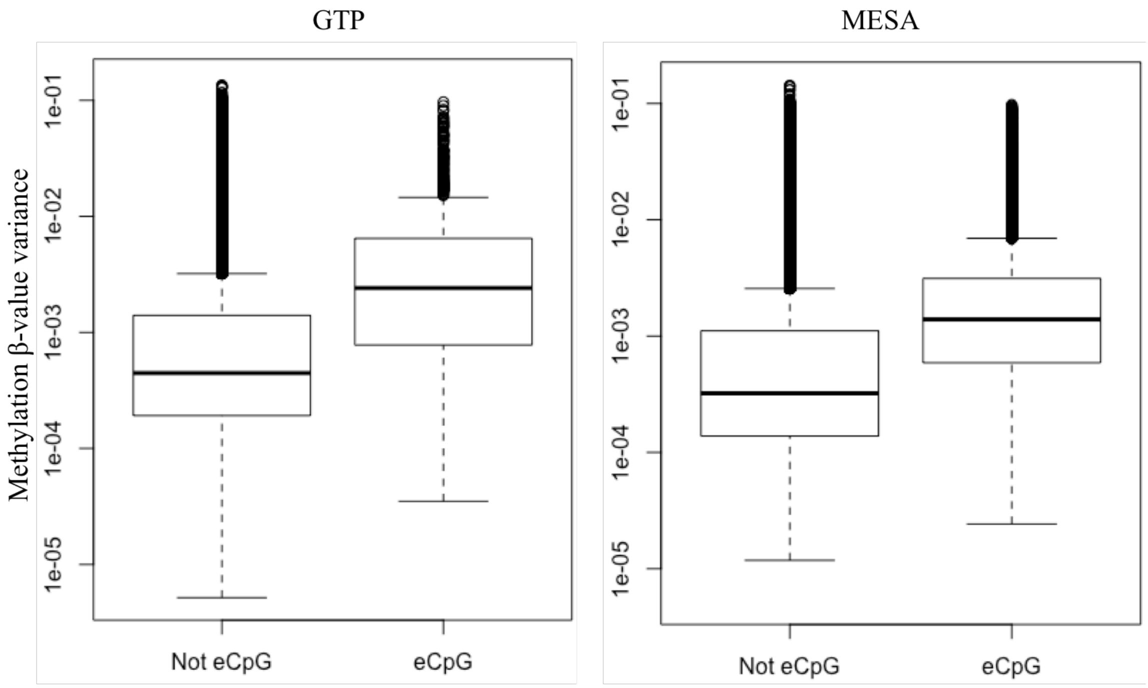


Figure S4-1. Distribution of DNA methylation in GTP and MESA. The expected bimodal distribution of DNA methylation β values is seen for all CpGs included in each study (broken lines). Genome-wide significant eCpGs from both studies tended to be intermediately methylated. eCpGs in GTP (light grey), were more heavily methylated and eCpG in MESA (dark grey) were more sparsely methylated, although both methylated and unmethylated fractions were present.


GTP

MESA

Mean: 1.6e-03* 6.4e-03* 1.7e-03* 3.6e-03*

*P<2.2e-16

Figure S4-2. Distribution of methylation beta-value variances, plotted on a log scale. Variances for either expression-associated or not-expression-associated CpGs were calculated across samples for GTP and MESA. The results indicate that eCpGs have more variable beta-values across samples.

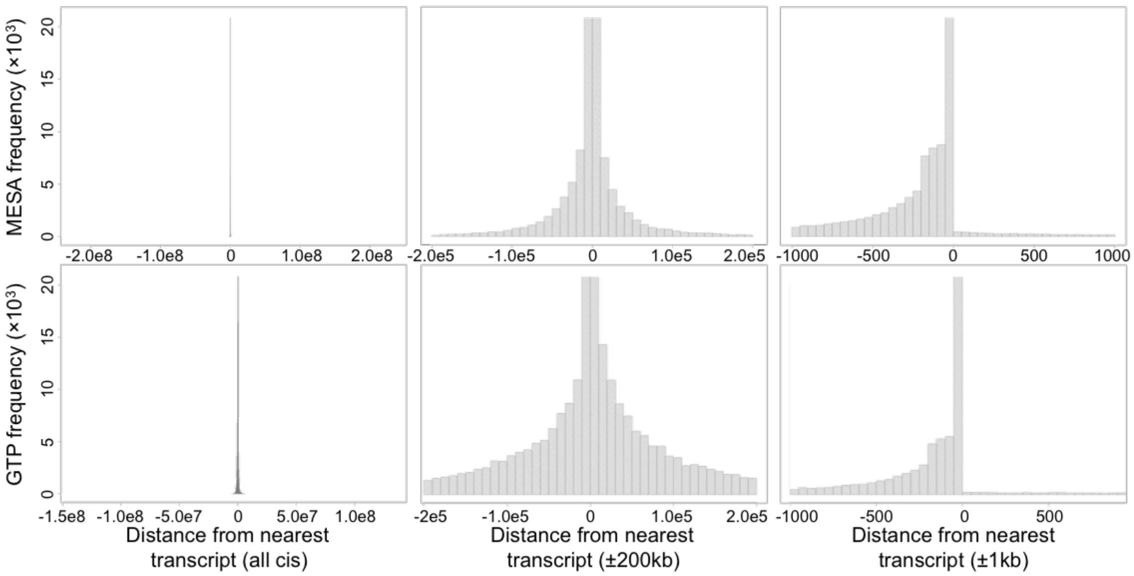
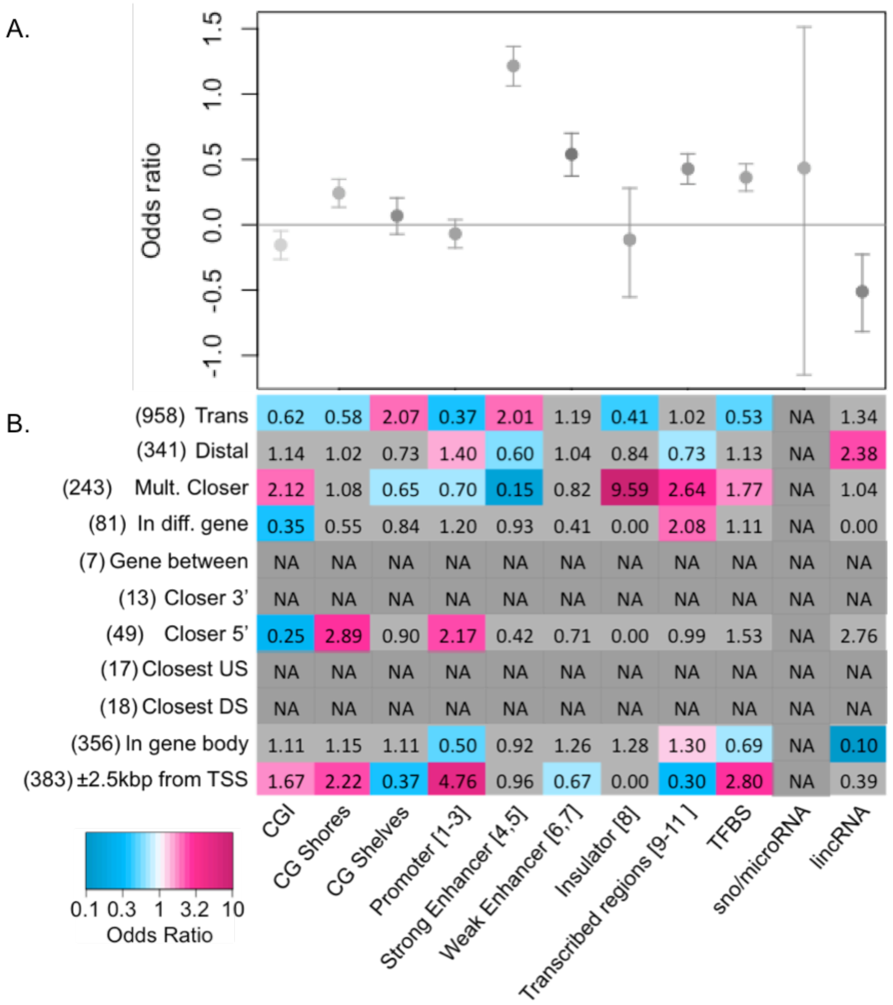


Figure S4-3. Distances between each CpG and the closest TSS for all CpG probes and expression probes included in GTP and MESA. Each panel contains the same data, focused on three different ranges of distances. The distribution of CpG-TSS distances for each array is similar to the distribution seen for eCpGs found in each study



	CGI	CG Shores	CG Shelves	Promoter [1-3]	Strong Enhancer [4,5]	Weak Enhancer [6,7]	Insulator [8]	Transcribed regions [9-11]	TFBS	sno/microRNA	lincRNA
(958) Trans	0.62	0.58	2.07	0.37	2.01	1.19	0.41	1.02	0.53	NA	1.34
(341) Distal	1.14	1.02	0.73	1.40	0.60	1.04	0.84	0.73	1.13	NA	2.38
(243) Mult. Closer	2.12	1.08	0.65	0.70	0.15	0.82	9.59	2.64	1.77	NA	1.04
(81) In diff. gene	0.35	0.55	0.84	1.20	0.93	0.41	0.00	2.08	1.11	NA	0.00
(7) Gene between	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
(13) Closer 3’	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
(49) Closer 5’	0.25	2.89	0.90	2.17	0.42	0.71	0.00	0.99	1.53	NA	2.76
(17) Closest US	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
(18) Closest DS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
(356) In gene body	1.11	1.15	1.11	0.50	0.92	1.26	1.28	1.30	0.69	NA	0.10
(383) ±2.5kbp from TSS	1.67	2.22	0.37	4.76	0.96	0.67	0.00	0.30	2.80	NA	0.39

Figure S4-4. Odds ratios for enrichment of chromatin features among CpGs. A) Enrichment of eCpGs in the listed chromatin features, among all CpGs tested in GTP (N=472,199). **B)** Enrichment of eCpGs in the listed chromatin features, among genome-wide significant eCpGs in GTP (N=2,466). Blue indicates significant depletion and red, significant enrichment. Light gray cells were not significant and ORs in dark gray cells could not be estimated due to low counts. Bracketed numbers in the chromatin features indicate the ChromHMM state. Numbers in parentheses indicate the number of eCpGs in the category. **Definitions: Bottom.** “CGI” are CpG islands. “TFBS” is transcription factor binding site.

Supplemental methods

Study Populations

The Grady Trauma Project (GTP). GTP is a prospective study of stress-related outcomes in which participants were recruited from the waiting rooms of Grady Memorial Hospital's General Practice or Obstetrics and Gynecology departments. Participants are from an inner city population with higher than average rates of trauma exposure and post-traumatic stress disorder, but are representative of this population in that they are not specifically ascertained for presence of disease or trauma. Exclusion criteria included mental retardation, active psychosis, or the inability to give written and verbal informed consent. The Institutional Review Boards of Emory University School of Medicine and Grady Memorial Hospital approved all procedures in this study. Genome-wide DNA methylation and gene expression measurements were generated for 333 human blood samples collected in Atlanta, Georgia as part of the Grady Trauma Project [50]. GTP participants included in this study range between 18 and 78 years old. 76% are female and all are African-American, based on self report and confirmation via principal component analysis of genotype data [51].

For gene expression analysis, whole blood was collected in Tempus RNA tubes. All whole genome expression profiles were generated at the Max-Planck Institute. RNA was isolated using the Versagene kit (Gentra Systems, Minneapolis, U.S.A.) and quantified using the Nanophotometer (Implen, München, Germany). Quality checks were performed on the Agilent Bioanalyzer. 250 nanograms of RNA were reverse transcribed to cDNA, converted to cRNA and biotin-labeled using the Ambion kit (AMIL1791, Applied Biosystems). 750 nanograms of cRNA were hybridized to Illumina HT-12 v3.0

or v4.0 arrays (Illumina, San Diego, California, U.S.A) and incubated for 16 hours at 55ºC. Arrays were then washed, stained with Cy3 labeled streptavidin, dried and scanned on the Illumina BeadScan confocal laser scanner. Expression values were normalized using the variance stabilizing transformation. 13,933 transcripts from the v3.0 and v4.0 arrays and were significantly expressed above background levels (detection $P<0.01$) in at least 5% of subjects, and were used in further analysis.

DNA was extracted from whole blood at the Max-Planck Institute in Munich using the Gentra Puregene Kit (Qiagen), to assay CpG methylation. Genomic DNA was bisulfite converted using the Zymo EZ-96 DNA Methylation Kit (Zymo Research) and applied to the Illumina HumanMethylation450 BeadChip, with hybridization and processing performed according to the instructions of the manufacturer. Methylated (M) and unmethylated (U) signals were collected for all CpG sites on the array and quantile normalized across all samples. β-values for each individual at each CpG site were calculated as the total methylated signal divided by the total signal (M/M+U). Data points with 1) a detection p-value greater than 0.001 or 2) a combined signal less than 25% of the total median signal and less than both the median unmethylated and median methylated signal were set to missing. CpG sites with a missingness rate above 10% were removed from analysis. Individual samples were removed from analysis if they were outliers in a hierarchical clustering analysis or had 1) a mean total signal less than half of the median overall mean signal or 2000 arbitrary units, or 2) a missingness rate above 5%.

Multi-ethnic Study of Atherosclerosis (MESA). MESA DNA methylation and gene expression data were collected from the Gene Expression Omnibus (GSE56047). The following summarizes the methods detailed in Liu et al. (2013).

MESA is a study designed to examine cardiovascular disease. The MESA Epigenomics and Transcriptomics Study investigates gene expression regulatory methylation sites in humans by examining the association between CpG methylation and gene expression in purified human monocytes from a large study population. Genome-wide DNA methylation and gene expression measurements were generated for purified monocytes from 1,202 MESA participants. These MESA participants were chosen randomly from samples collected between April 2010 and February 2012 from MESA field centers in Baltimore, MD; Forsyth County, NC; New York, NY; and St Paul, MN. Participants range in age from 55 – 94 years old, are 51% female, and self identified as Caucasian (47%), African American (21%), or Hispanic (32%).

Monocytes were isolated from PBMCs with anti-CD14 coated magnetic beads. DNA and RNA were isolated from monocyte samples simultaneously. DNA and RNA purity were assessed spectrophotometrically and RNA QC testing was performed using the Agilent 2100 Bioanalyzer with RNA 6000 Nano chips (Agilent Techonology, Inc., Santa Clara, CA) according to the manufacturer's instructions. Samples with RIN (RNA Integrity) scores > 9.0 were applied to global gene expression microarrays.

Genome-wide expression analysis was performed via the Illumina HumanHT-12 v4 Expression BeadChip and the Illumina Bead Array Reader, following the Illumina expression protocol. RNA was reverse transcribed and amplified with the Illumina

TotalPrep-96 RNA Amplification Kit (Ambion/Applied Biosystems, Darmstadt, Germany). The resulting cRNA was hybridized to a BeadChip (CITATION).

Expression data were corrected for local background in Illumina's proprietary software GenomeStudio and negative controls were used to compute detection P-values. Normal-exponential convolution model analysis was used to estimate non-negative signal. All probes and samples were quantile normalized, offset and $\log_2$ transformed [23]. Expression probes with detection p-values >0.01 in at least 5% of samples were removed.

Bead-level methylation data were summarized in and collected from GenomeStudio. Methylation data were smooth quantile normalization (to adjust for color bias) normalized, background corrected (by subtracting the median intensity of the negative control probes), and quantile normalized across all samples with the R package, lumi. [23]. The final methylation value for each methylation probe was computed as the *M*-value [*M* is logit(beta-value)]. We collected these values from GEO, transformed the *M*-values to β-values and performed additional QC (concerning missingness and low signal intensity; as in the GTP methylation data).

Transcript annotation. The Illumina HT12 probe locations were provided a re-annotation of the array (Bioconductor packages illuminaHumanv3.db and illuminaHumanv4.db; Barbosa-Morais et al., 2010). Probes annotated as "bad" or "'no match" were removed from analysis. Each probe was assigned multiple genomic locations, based on the "GenomicLocation", "SecondMatches", and "OtherGenomicMatches" listed in the re-annotation. Refseq and Ensembl transcript and exon intervals for the HG19 build were collected from the UCSC table browser. When exons for the same gene overlapped, the start and end points yielding the largest interval were taken to make a representative or average exon. Refseq and Ensembl gene information was annotated to each expression probe ID (using R Bioconductor package GenomicRanges; [52] if the probe interval overlapped the exon interval by more than 25 bp. Where more than one gene exon overlapped the expression probe, poorly characterized genes (gene names beginning with KIAA, FLJ or LOC and those containing “orf”) and read-through transcripts were removed from the duplicates. The remaining duplicate gene names were compiled for that probe ID and the transcript area taken as the largest interval formed by the overlapping transcripts. Similarly, if overlapping transcript areas (transcription start site to transcription end site, including introns) existed for the same gene name in the Refseq or Ensembl tables, the minimum TSS and maximum TES were taken for the gene entry. Refseq or Ensembl tables with representative transcript intervals are referred to as average Refseq and average Ensembl gene locations in the Materials and Methods section.

Assignment of probe location. For each eCpG-transcript pair, all probe locations (for one Probe ID) were compared to the eCpG location. Probe locations were prioritized in the following order:

1) The eCpG location fell within the gene annotated to one of the probe locations or up to 2,500 bp upstream of that gene’s TSS.
2) The eCpG was within 1 megabase of the TSS of the gene annotated to one the probe locations.
3) The eCpG was on the same chromosome as one the probe locations.

If more than one probe location fell into the highest priority group, they were further filtered by the following criteria until one probe location was chosen: locations annotated to a gene are preferred, locations marked as GenomicLocation > SecondMatches > OtherGenomicMatches, locations in which the eCpG was closest to the TSS of the annotated gene were preferred. If all probe locations were on different chromosomes than the eCpG, probe locations were chosen as follows: locations annotated to a gene are preferred, locations marked as GenomicLocation > SecondMatches > OtherGenomicMatches.

References

1. Richardson BC. Role of DNA methylation in the regulation of cell function: autoimmunity, aging and cancer. J Nutr. 2002;132 8 Suppl:2401S–2405S.

2. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.

3. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144:327–39.

4. Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. Trends Genet TIG. 2007;23:413–8.

5. Florath I, Butterbach K, Heiss J, Bewerunge-Hudler M, Zhang Y, Schöttker B, et al. Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. Diabetologia. 2016;59:130–8.

6. Freeman JR, Chu S, Hsu T, Huang Y-T. Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. Oncotarget. 2016;7:69579–91.

7. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9:436–47.

8. Julià A, Absher D, López-Lasanta M, Palau N, Pluma A, Waite Jones L, et al. Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. Hum Mol Genet. 2017;26:2803–11.

9. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Kheradpour P, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

10. Ligthart S, Steenaard RV, Peters MJ, Meurs JBJ van, Sijbrands EJG, Uitterlinden AG, et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. Diabetologia. 2016;59:998–1006.

11. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2016.

12. Nagarajan RP, Zhang B, Bell RJA, Johnson BE, Olshen AB, Sundaram V, et al. Recurrent epimutations activate gene body promoters in primary glioblastoma. Genome Res. 2014;24:761.

13. Ventham NT, Kennedy NA, Adams AT, Kalla R, Heath S, O'Leary KR, et al. Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. Nat Commun. 2016;7:13507.

14. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6.

15. Bell CG. Epigenome-Wide Association Studies: Potential Insights into Human Disease. In: Naumova AK, Greenwood CMT, editors. Epigenetics and Complex Traits. Springer New York; 2013. p. 287–317. http://link.springer.com.proxy.library.emory.edu/chapter/10.1007/978-1-4614-8078-5_13. Accessed 28 Apr 2014.

16. Harris RA, Nagy-Szakal D, Pedersen N, Opekun A, Bronsky J, Munkholm P, et al. Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. Inflamm Bowel Dis. 2012;18:2334–41.

17. Lin Z, Hegarty J, Cappel J, Yu W, Chen X, Faber P, et al. Identification of disease-associated DNA methylation in intestinal tissues from patients with inflammatory bowel disease. Clin Genet. 2011;80:59–67.

18. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14:1–14.

19. Eijk KR van, Jong S de, Boks MP, Langeveld T, Colas F, Veldink JH, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics. 2012;13:636.

20. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12:R10.

21. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. Am J Hum Genet. 2013;93:876–90.

22. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15:R37.

23. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, et al. Methylomics of gene expression in human monocytes. Hum Mol Genet. 2013;22:5065–74.

24. Joo JWJ, Sul JH, Han B, Ye C, Eskin E. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. Genome Biol. 2014;15:R61.
25. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JFJ, Ritchie ME, Lynch AG, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res. 2010;38:e17–e17.
26. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42 Database issue:D756-763.
27. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
29. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.
30. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
31. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinforma Oxf Engl. 2007;23:257–8.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 57:289–300.

33. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PloS One. 2012;7:e41361.

34. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32 Database issue:D493-496.

35. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466:253–7.

36. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. Cancer Cell. 2014;26:577–90.

37. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015;520:243–7.

38. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. Oncotarget. 2012;3:462–74.

39. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife. 2013;2:e00523.

40. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform. 2016;17:980–95.

41. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol Cell. 2017;66:711–720.e3.

42. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet. 2012;13:271–82.

43. Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. Nat Rev Cancer. 2015;15:321–33.

44. Long X-R, He Y, Huang C, Li J. MicroRNA-148a is silenced by hypermethylation and interacts with DNA methyltransferase 1 in hepatocellular carcinogenesis. Int J Oncol. 2014;44:1915–22.

45. Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet. 2015;16:71–84.

46. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

47. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23:555–67.

48. Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009;27:361–8.

49. Yang X, Shao X, Gao L, Zhang S. Systematic DNA methylation analysis of multiple cell lines reveals common and specific patterns within and across tissues of origin. Hum Mol Genet. 2015;24:4374–84.

50. Gillespie CF, Bradley B, Mercer K, Smith AK, Conneely K, Gapen M, et al. Trauma Exposure and Stress-Related Disorders in Inner City Primary Care Patients. Gen Hosp Psychiatry. 2009;31:505–14.

51. Kilaru V, Iyer SV, Almli LM, Stevens JS, Lori A, Jovanovic T, et al. Genome-wide gene-based analysis suggests an association between Neuroligin 1 (NLGN1) and post-traumatic stress disorder. Transl Psychiatry. 2016;6:e820.

52. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol. 2013;9:e1003118.

CHAPTER V. Discussion

Common challenges emerge from different epigenetic studies

DNA methylation is central in the study of human epigenetics. This work addresses two key aspects of DNA methylation: how patterns change in response to environment and how those changes have downstream biological effects. Chapters II and III provide evidence that environmental exposures and time play an important role in the dynamics of DNA methylation over a person's lifetime. The results presented in Chapter III suggest that exposures, like irradiation, can leave a lasting imprint on the DNA methylome, and that imprint in turn may account for some of the risk associated with radiation exposure. While this information is surely important to keep in mind as humans set a course for space, it also has terrestrial applications in radiation oncology and other modes of radiation exposure. Chapter IV proposes that DNA methylation can have far-reaching associations in gene expression, potentially through previously undescribed channels such as non-coding RNAs. The results of this study suggest that the role of DNA methylation in the genome might not be as straightforward as previously thought.

However different these two studies of DNA methylation might be, common themes emerge and highlight key areas where progress can be made in epigenomics, both in methodology for analysis and in interpretation of results. As is representative of all epigenomic studies, the conclusions that could be drawn in Chapters III and IV were limited by potential confounding due to cell type heterogeneity and other unmeasured factors, but this was partially addressable through the use of appropriate statistical methodologies. In terms of results, both studies yielded large sets of associated CpG sites to be interpreted, and the leveraging of several sources of –omics data broadened the interpretability of results in both studies. Results like those presented in Chapter IV have the potential to similarly enhance interpretation of results in future studies, and represent a resource to the field and a source of novel insights into the regulatory role of DNA methylation. Lastly, in addition to new insights into the human epigenome, both studies provide a wealth of new hypotheses that can and should be corroborated, *ex silico*. The remaining sections discuss each of the issues raised above, in turn, followed by a brief reflection on how my doctoral work fits within the field of epigenomics.

Overcoming statistical confounding

Statistical confounding in EWAS is a daunting hurdle. Common sources of confounding are cell-type proportion and age [1]. Although the ICE framework [2] was effective in correcting for unknown confounding in Chapter IV, controlling statistical inflation was more far more challenging in GTP than in MESA. The Houseman method [3], which is currently a gold-standard bioinformatic approach for inferring and accounting for cell-type proportion differences between samples, could not ameliorate the confounding in the model of GTP associations. While one can never be certain of the sources of confounding, it is tempting to postulate that the root of the problem was the mixed cell type in GTP's whole blood samples. For EWAS to be successful, samples that are of a single cell type are ideal. However, while this ideal is difficult to meet in any study of human subjects due to the heterogeneity of non-invasively collectible tissues (such as whole blood) and the costs involved in flow sorting, it seems nearly impossible in cancer studies, where tumors can be highly heterogeneous [4]. It is precisely because sources of statistical confounding can be difficult to identify, that regression frameworks that account for unknown confounding, like ICE, should be considered for use in EWAS studies.

Although it is generally accepted that DNA methylation patterns can and do change with time in individuals, Chapter III also shows that methylation patterns in cultured cells change drastically over time. This sort of confounding is relatively easy to control in one experiment or laboratory, as time in culture and passage number are known values. The harder, and perhaps more important question to answer is how comparable cell cultures from different studies or laboratories are, epigenetically. Genomics and epigenomics researchers often rely on publicly available data, which is often derived from cell cultures. Care must be taken to ensure that those data are appropriate to extrapolate to samples from other studies of a similar cell type.

Combining –omics for improved interpretation

Chapters III and IV both combined results from other genome-wide, -omics studies. Both studies used the ChromHMM data from ENCODE [5] to assess annotated chromatin features (e.g., promoter, enhancer, heterochromatin) among CpGs associated with radiation exposure and among eCpGs of various type. They both also used a map of CpG islands available through the UCSC genome browser. DNA methylation data for two lung cancer studies, available through The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/), were leveraged in Chapter III to reveal parallels between radiation-associated CpGs and epigenetic patterns in human lung cancer. Finally, Chapter IV utilized gene annotations [6, 7], transcription factor binding site maps, non-coding RNA maps [8, 9] (http://genome.ucsc.edu/) and HiC interaction maps [10] to reveal both direct and indirect ways that DNA methylation can associate with gene expression, genome-wide. These resources are all publicly available and have the potential to vastly improve functional interpretation for many types of studies.

While irradiation resulted in interesting epigenetic patterns in annotated functional categories at the chromatin-level, no clear gene-level associations could be inferred between radiation treatment and DNA methylation in Chapter III. As discussed in Chapter IV, it is not uncommon for EWAS to report genome-wide, moderately significant results. The DNA methylation patterns associated with the expression of each transcript in Chapter IV were often not near the TSS or gene body (non-canonical) and contained a mix of cis, distal and trans associations. Unfortunately, the epigenetic patterns of irradiation and gene expression identified in the two studies originated from two different tissues (lung vs. blood cells), making them incomparable across studies. However, had the two studies been conducted in similar cell types, the epigenomic patterns of transcription for each gene (Chapter IV) could have been compared to the EWAS results for irradiation (Chapter III) and could have potentially provided more gene-level insight. Going forward, studies of eCpGs in varying tissues could provide the epigenomics field with an invaluable tool for EWAS interpretation.

As the field of epigenomics moves forward, studies and analytical methods will continue to employ larger sample sizes and more refined analytical methods, increasing statistical power. However, larger sample sizes are often balanced by a larger multiple testing burden, as technology allows for the collection of methylation data at increasing density. Where researchers were once assaying 27k [11] and then 450k CpGs [12], they are now assaying closer to 850k CpGs or utilizing whole genome bisulfite sequencing. While increasingly dense studies of the epigenome require an increasingly stringent multiple testing burden, they also provide greater opportunities to understand the role of DNA methylation in health and disease. However, as revealed in Chapter IV, the regulatory role of specific CpG sites is not always straightforward, and asking more questions without the means for relevant interpretation is unlikely to advance current understanding. Resources like those listed above will be key to meaningful interpretation of results.

Combining bench and population science

Statistical and bioinformatic tools will continue to be fundamental tools for finding and interpreting associations of DNA methylation and a phenotype of interest, but they are currently (and will likely continue to be) hampered by reverse causation [1]. Because DNA methylation can change in response to environment, cross sectional and case-control studies of disease cannot easily differentiate between a CpG at which methylation plays a causal role in disease and a CpG at which methylation is a byproduct of processes related to disease.

In order to fully understand the role of DNA methylation in the genome, *in silico* observations should be confirmed experimentally. This has proven more difficult for studies of epigenetics than genetics, where protocols exist for experimentally modifying DNA sequence. Listed below is a brief evaluation of a few experimental tools available to confirm the suspected roles of DNA methylation.

The DNA demethylating drug, 5-aza-2′-deoxycytidine, is a cytidine analog that is incorporated into replicating DNA and will covalently bind to the catalytic sites of all three biologically active DNMTs, triggering their degradation. Treatment with 5-aza-2′-deoxycytidine results in genome-wide loss of DNA methylation and has been shown to activate transcription of epigenetically silenced genes in cancer (reviewed in [13]). However, this approach is non-specific and only targets a subset of genes [14–16]. Chapter IV evidences a phenomenon in which CpG methylation from multiple locations of the epigenome can influence gene expression. Treatment with a non-specific demethylating drug, like 5-aza-2′-deoxycytidine would not allow researchers to determine which specific methylation changes were driving changes in gene expression or which were responsible for the phenotype of interest.

A more specific approach to testing bioinformatic findings *in situ* is through a reporter assay. Proposed regulatory elements can be inserted into vectors containing a CpG-free promoter and reporter gene (like luciferase [17–19]). Methylation can be induced in only the inserted element and the complete plasmid transfected into cultured human cells for assessment of luciferase activity. This assay is far more specific than drug treatment, but disregards native context, which may or may not be important. For instance, CTCF-bound insulator sequences can be vital for promoter-enhancer interactions [20], but would not be present in a reporter assay.

Finally, in recent years, genome-editing tools like zinc-finger nucleases [21], transcription activator-like effector nucleases (TALENs) [22] and the CRISPR-Cas9 system [23] have allowed researchers unparalleled ability to modify the genetic code. More recently, these same systems have been exploited to combine sequence specificity with the catalytic activity of epigenome-modifying enzymes, like histone methyltransferases and deacetylases, DNA methyltransferases and Ten-Eleven Translocation 2 [24–30]. Fusion proteins have allowed these systems to be used for epigenome editing. Although these approaches are still in their infancy, they represent a promising leap in epigenetics. These systems have the potential to be far more specific than epigenetic drugs, while also allowing researchers to test epigenetic hypotheses in the native context of the genome.

Conclusion

DNA methylation is a dynamic mark in the epigenome with links to regulation of gene expression, age, environmental exposures and various diseases. This body of work has shed light on some of the ways in which epigenetic modifications bridge the gap between the genetic code and the phenotypic variation seen in every individual. It has shown that an individual’s environment can induce lasting epigenetic changes. It has also shown that these epigenetic changes can associate with changes to gene expression in unexpected ways. The conclusions drawn from this work are broad. Clearly, epigenetic changes must be considered, in addition to genetic mutations, with regard to space-travel and more broadly, radiation exposure. Environmental exposures will likely prove an important stochastic source of epigenetic variation, in general. Moreover, the function and role of DNA methylation in gene expression has been too narrowly defined. To draw accurate and meaningful conclusions from epigenomic studies, it is imperative that we understand the regulatory effects of changes to the epigenome. Finally, this work has highlighted key areas in which epigenomic studies will have to improve as the field moves forward. Epigenomics is key to understanding how the genetic code is translated into life and to understanding how each life is shaped. Perhaps the key to elucidating the results generated by epigenomic studies will be to work smarter, as well as larger.

References

1. Birney E, Smith GD, Greally JM. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. PLOS Genet. 2016;12:e1006105.

2. Joo JWJ, Sul JH, Han B, Ye C, Eskin E. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. Genome Biol. 2014;15:R61.

3. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

4. Easwaran H, Baylin SB. Epigenetic Abnormalities in Cancer Find a "Home on the Range." Cancer Cell. 2013;23:1–3.

5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

6. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42 Database issue:D756-763.

7. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.

8. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25:1915–27.

9. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34 Database issue:D140–4.

10. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

11. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1:177–200.

12. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288–95.

13. Ahuja N, Sharma AR, Baylin SB. Epigenetic Therapeutics: A New Weapon in the War Against Cancer. Annu Rev Med. 2016;67:73–89.

14. Hagemann S, Heil O, Lyko F, Brueckner B. Azacytidine and Decitabine Induce Gene-Specific and Non-Random DNA Demethylation in Human Cancer Cell Lines. PLOS ONE. 2011;6:e17388.

15. Flotho C, Claus R, Batz C, Schneider M, Sandrock I, Ihde S, et al. The DNA methyltransferase inhibitors azacitidine, decitabine and zebularine exert differential effects on cancer gene expression in acute myeloid leukemia cells. Leukemia. 2009;23:1019–28.

16. Stresemann C, Lyko F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. Int J Cancer. 2008;123:8–13.

17. Kennedy A, Schmidt EM, Cribbs AP, Penn H, Amjadi P, Syed K, et al. A novel upstream enhancer of FOXP3, sensitive to methylation-induced silencing, exhibits dysregulated methylation in rheumatoid arthritis Treg cells. Eur J Immunol. 2014;44:2968–78.

18. Ehrlich KC, Paterson HL, Lacey M, Ehrlich M. DNA Hypomethylation in Intragenic and Intergenic Enhancer Chromatin of Muscle-Specific Genes Usually Correlates with their Expression. Yale J Biol Med. 2016;89:441–55.

19. Ishihara K, Nakamoto M, Nakao M. DNA methylation-independent removable insulator controls chromatin remodeling at the HOXA locus via retinoic acid signaling. Hum Mol Genet. 2016;25:5383–94.

20. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform. 2016;17:980–95.

21. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. Nat Rev Genet. 2010;11:636–46.

22. Doyle EL, Stoddard BL, Voytas DF, Bogdanove AJ. TAL effectors: highly adaptable phytobacterial virulence factors and readily engineered DNA-targeting proteins. Trends Cell Biol. 2013;23:390–8.

23. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013;8:2281–308.

24. Nunna S, Reinhardt R, Ragozin S, Jeltsch A. Targeted methylation of the epithelial cell adhesion molecule (EpCAM) promoter to silence its expression in ovarian cancer cells. PloS One. 2014;9:e87703.

25. Vojta A, Dobrinić P, Tadić V, Bočkor L, Korać P, Julg B, et al. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res. 2016;44:5615–28.

26. Rivenbark AG, Stolzenburg S, Beltran AS, Yuan X, Rots MG, Strahl BD, et al. Epigenetic reprogramming of cancer cells via targeted DNA methylation. Epigenetics. 2012;7:350–60.

27. Chen H, Kazemier HG, de Groote ML, Ruiters MHJ, Xu G-L, Rots MG. Induced DNA demethylation by targeting Ten-Eleven Translocation 2 to the human ICAM-1 promoter. Nucleic Acids Res. 2014;42:1563–74.

28. Falahi F, Huisman C, Kazemier HG, Vlies P van der, Kok K, Hospers GAP, et al. Towards Sustained Silencing of HER2/neu in Cancer By Epigenetic Editing. Mol Cancer Res. 2013;11:1029–39.

29. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33:510–7.

30. Mendenhall EM, Williamson KE, Reyon D, Zou JY, Ram O, Joung JK, et al. Locus-specific editing of histone modifications at endogenous enhancers. Nat Biotechnol. 2013;31:1133–6.