Distribution agreement

Signature:

Joshua SK Bell Date

Resetting the Cancer Transcriptome with Epigenetic Therapy

Joshua S.K. Bell

Doctor of Philosophy

Graduate Division of Biological and Biomedical Science

Genetics and Molecular Biology

Paula M. Vertino, Ph.D

Advisor

Xiaodong Cheng, Ph.D.

Committee Member

Jeremy Boss, Ph.D.

Committee Member

Victor Corces, Ph.D.

Committee Member

Carlos Moreno, Ph.D.

Committee MemberAccepted:

Lisa A. Tedesco, Ph.D.

Dean of the James T. Laney School of Graduate Studies

Date

Resetting the Cancer Transcriptome with Epigenetic Therapy

By
Joshua S.K. Bell
B.S., University of Georgia, 2008

Advisor: Paula Vertino, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology
2017

Abstract

DNA methylation is a key regulator of transcription in mammals, and aberrant methylation changes drive tumorigenesis. Epigenetic therapy inhibiting DNA methylation decreases the growth rate and invasive potential of solid tumor cells, and is a standard treatment for certain hematological cancers. However, many questions remain surrounding the exact mechanism of this epigenetic reprogramming, and the role of methylation in distinct genomic compartments. The human genome is a heavily methylated, CpG-depleted, dominantly heterochromatic terrain disrupted by CpG Islands (CGI), essential hypomethylated CpG dense regulatory elements, associated with most human promoters. CGI are characterized by high levels of transcriptional initiation, but transcription into the gene body is limited by promoter-proximal RNA Polymerase II (Pol II) pausing. Here, we document an additional Pol II pausing step at CGI boundaries known as distal pausing, and link the location and degree of pausing to GC-skew. Many 'orphan' CGI are also found far from any annotated transcript, and we establish most such CGI are highly active enhancers. We also examine the nascent transcriptome of cancer cells using Precision Run-on Sequencing (ProSeq) during epigenetic therapy to document that DNA hypomethylation results in pervasive changes including the down-regulation of oncogenes due to loss of gene body methylation, reactivation of genes silenced by promoter methylation, reactivation of enhancers hypermethylated in breast cancer, and induction of repetitive elements. Critically, many of these transcripts remain highly paused or are unstable and thus have evaded detection by RNA-Seq. Analyzing remethylation kinetics following therapy, we find that normal methylation returns quickly, while aberrant cancer-related hypermethylation is largely forgotten. These findings suggest new mechanisms by which epigenetic therapy reprograms cancer cells and asserts novel roles for DNA methylation in regulating transcription more broadly.

Epigenetic Therapy Resets the Cancer Transcriptome

By
Joshua S.K. Bell
B.S., University of Georgia, 2008

Advisor: Paula Vertino, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology
2017

Table of Contents

Chapter I: Introduction

Chapter II: GC-Skew Defines Distinct RNA Polymerase Pause Sites in CpG Island Promoters

Chapter III: Factors Affects the Persistence of Drug-Induced Reprogramming of the Cancer Methylome

Chapter IV: Orphan CpG Islands Define a Class of Highly Active Enhancers

Chapter V: Therapeutic Inhibition of DNA Methylation Resets the Cancer Transcriptome

Chapter VI: Discussion

Chapter I: Introduction

Epigenetics

Conrad Waddington originally coined the term 'epigenetics' in 1942[1] to describe the processes mediating 'epigenesis', the theory dating from Aristotle that organisms develop progressively by differentiation rather than 'preformationism',[2] the once commonly-held view that organisms develop from miniature versions of the adult form. Waddington had little idea of the molecular mechanisms behind development, with the fundamental demonstration that DNA was the genetic material not coming until 1944 [3]. Yet, chromosomes had been identified in 1879 by Walther Fleming, and following his seminal studies in *Drosophila,* Thomas Hunt Morgan in 1911 mapped specific genes to the X-chromosome. Indeed, some of the first hints that genetics alone were insufficient to account for phenotype came from so called 'eversporting' translocation mutations in *Drosophila*, where preservation of all chromosomal material, but in different rearrangements, resulted in distinct phenotypes, an epigenetic phenomenon now referred to as position effect variegation where neighboring chromatin spreads to silence or activate nearby genes.

Yet even with limited mechanistic knowledge, Waddington's theory of an 'epigenetic landscape', essentially that as cells develop they acquire distinct epigenetic states, has become dogma in biology. We know now that this epigenetic state encompasses an array of mitotically heritable molecular modifications to chromatin, most prominently DNA methylation and histone modifications and variants. It is clear that programmed epigenetic changes during development result in the expression of distinct gene sets that underlie the morphological and phenotypic changes to cells embryologists have long studied. Furthermore, we know now that aberrant epigenetic changes play a crucial role in human disease, especially in oncogenesis and cancer progression.

DNA Methylation & CpG Islands

One of the earliest examples of a clear role for epigenetics in gene regulation comes from X-chromosome inactivation. In female placental mammals, it was demonstrated at least at early as the 1950s that one X-chromosome is randomly epigenetically silenced early in development[4,5] and in 1975 it was proposed that DNA methylation mediated this silencing[6,7], a fact confirmed by dozens of studies since[8]. Indeed, DNA methylation is now perhaps the best-studied chromatin mark, with essential roles in development, gene silencing, aging, and carcinogenesis. DNA methylation occurs at the 5-carbon of cytosine, and is a common mark conserved through bacteria, fungi, plants, and animals[9].

In mammals, CpG dinucleotides in particular are targeted for DNA methylation during development unless actively protected by H3K4 methylation (discussed below), a mark found predominantly at enhancers and promoters and associated with transcriptional initiation[10–12] (although CpA methylation does occur in embryonic stem cells and certain neurons[13].) Responsible for this CpG targeting are the DNA methyltransferase enzymes DNMT1, 3A & 3B. Each is essential for development[14], and intriguingly only embryonic stem cells, but not differentiated or cancer cells, can survive a complete absence of DNA methylation. DNMT1 is specifically targeted to hemimethylated DNA during mitosis, ensuring faithful epigenetic maintenance, while DNMT3A &B have traditionally been known as the *de novo* methyltransferases, targeted by other factors to unmethylated DNA, although each is also important for maintenance methylation[15] .

DNA methylation is a substrate for hydroxymethylation (5hmC) by the TET family of enzymes. 5hmC levels are highest in embyronic stem cells and certain kinds of neurons, but can be detected at low levels in most human cell types. 5hmC can serve as an intermediate in

DNA demethylation, but also has functions in its own right. 5hmC can participate in both active and passive demethylation: in the former, TET enzymes further enzymatically oxidize 5hmC to 5-carboxy- and 5-formyl-cytosine, which can be excised by DNA repair machinery and replaced with unmodified cytosine. DNMT1 has lower fidelity for 5hmC than 5mC and so during mitosis, it is thought that 5hmC can passively impair transmission of methylation to daughter cells. Hydroxymethylation is found at extremely high levels in super enhancers, and may also serve a functional role itself, either to buffer the binding of 5hmC-sensitive TFs or to specifically recruit 5hmC readers, many of which have recently been identified[16].

Unlike unmodifed cytosine, methylated cytosine is prone to spontaneous deamination to thymine, which has led to a loss of CpG dinucleotides in mammalian genomes over evolutionary time. This process has led to unmethylated regions of high CpG content known as CpG Islands (CGI), with apparent functional importance given their selection to remain unmethylated, in an otherwise heavily methylated genome.

CGI were originally isolated in the 1980s as a fraction of the genome digested by Hpa II, a restriction enzyme only capable of cleaving unmethylated DNA[17]. This fraction separates distinctly from the heavily methylated genome when run on an agarose gel, leading to the 'island' designation. In the decades since, it has become clear that CGI are critical regulatory elements in the genome. CGI, in addition to their hypomethylation and CpG density, are known for their transcriptional competence, and heightened euchromatic features[18] ,and are found at nearly two-thirds of human protein-coding promoters. Hypomethylation is generally thought to be an intrinsic feature of CGI, as removal or addition of specific residues is sufficient to render such genomic regions sensitive or resistant to DNA methylation[19,20]. CGI have canonically been studied for their role in promoting transcriptional initiation, and promoters containing CGI demonstrate broader expression and greater strength than CpG-poor promoters[21]. These properties are contingent on hypomethylation: hypermethylation of CGI generally recruits repressive histone deacetylases and chromatin remodelers[25] leading to silencing of the associated gene, often in a programmed developmental context[23–25] ,but is also a cardinal method by which cancer cells inactivate tumor-suppressor genes[26–28].

Unlike other promoters, CGI often lack TATA boxes and appear more dependent on their GC content for activity. Indeed, CGI promoters do not show extensive sequence conservation in mammals except for in CpG sites[18]. One reason for this may be that transcription factors (TFs) on average tend to bind more GC-rich DNA, and a number of TFs specifically bind CpG sites, often in a methylation-state dependent manner[29].Compared to other promoters, CGI exhibit much more intense, bidirectional transcriptional initiation when unmethylated. However, RNA Polymerase II pausing (discussed below) is a critical check on gene activation at CGI, and while hypomethylation is permissive to initiation, it does not guarantee gene expression.

In the early days of epigenetics, it was difficult to map marks to specific regions. For example, early studies of DNA methylation relied on restriction digest with methylation-sensitive enzymes, and histone modifications could only be visualized in bulk. Current techniques for measuring DNA methylation now rely primarily on bisulfite sequencing. Treatment with sodium bisulfite converts unmethylated cytosines to uracil, while methylated cytosines are protected. The advent of PCR allowed use of bisulfite conversion to map methylation at base and allele resolution, but modern techniques like whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS), which first isolates CpG-rich DNA with restriction enzymes, have enabled genome-wide analysis of DNA methylation. In this work, we utilize array-based techniques that distinguish converted cytosines based on hybridization to guanine- or adenine-containing probes. There have been several generations of such arrays, with the first assaying just a few hundred sites and the current Illumina Epic array measuring over 850 thousand CpGs at promoters, gene bodies, non-coding RNA loci, and enhancers, among other intergenic regions[30].

Histone Modifications

In cells, DNA is tightly packed and organized around proteins known as histones, in approximately 146bp units known as nucleosomes. In mammals, histones are found in octamers with two copies each of H2A, H2B, H3, and H4, each of which is subject to post-translational modifications including methylation, acetylation, phosphorylation and numerous other modifications at various residues. Since the 1960s it has been clear that such modifications influence gene expression[31]. These modifications can broadly be divided into activating and respressive marks, although many marks play more complicated roles, that collectively compose the 'histone code', a set of instructions to cells dictating gene expression. Intrinsic to the histone code are 'writers', 'readers', and 'erasers', which enzymatically modify histones to respectively add, recognize, or remove these modifications. Each unique mark has its own family of regulators. For example, most histone lysine methylases (writers) contain a SET domain, named for the founding members of the family in *Drosphila* (Su(var)3-9, Enhancer of zeste, and trithorax). And while DOT1L, the writer of H3K79me, possesses a distinct enzymatic domain, both families use S-adenosylmethionine (SAM) as a cofactor/methyl-donor. There are also two families of lysine demethylases (erasers), amine oxidases (like LSD1) and iron-dependent dioxigenases, which contain a jumonji-C domain (such as JMJD1C). Reading of lysine methylation can be accomplished by a slew of methyl-lysine-binding domains including PWWP domains, WD40 repeats, PHD fingers, ankyrin repeats, and members of the Royal superfamily: chromodomains, chromobarrels, Tudor domains, and MBT repeats, among others.

Among the most important modifications associated with active transcription is methylation of H3K4. Nearly a dozen distinct histone methylases and demethylases can act on H3K4 in different cell types and contexts, and perhaps dozens of readers recognize the different methylation states (me0-me3)[32]. Progressive H3K4me states are associated with greater levels of transcriptional initiation. H3K4me1 tends to be associated with enhancer elements (discussed below), which display modest transcription of enhancer RNAs (eRNA). H3K4me3 tends to be found at active promoters, especially promoter CGI. H3K4me2 is less studied, but is typically linked to the transition between H3K4me1 and H3K4me3.

Methylation of H3K4 is tightly linked to histone acetylation, which can occur at many lysine residues and is nearly always linked to transcriptional activity. Lysine acetylation neutralizes the positive charge of lysine residues, leading to decompaction of chromatin given the negative charge of DNA. Histone acetylation is written by scores of histone acetyltransferases (HATs), erased by histone deacetylases (HDACs), and read by many proteins, among the most well-studied of which contain bromodomains specifically dedicated to the task [33]. Bromodomains are found in many proteins including HATs, helicases, transcription factors, ATP-dependent chromatin remodellers, and histone methylases. Among the most well-studied acetylation marks are H3K9Ac and H3K27Ac, although HATs tend to exhibit broad activity and acetylation levels at multiple lysine residues is often linked. Both of these marks are found at high levels at promoter CGI, but are also associated with enhancers. Histone acetylation tends to be uncommon in gene bodies, perhaps to prevent spurious initiation.

Methylation of H3K36 is also found at actively transcribed genes, but facilitates transcription through gene bodies, and actually appears to repress transcriptional initiation. H3K36me3 is preferentially found at highly expressed genes, and is written by SETD2 (dependent on lower methylation states written by at least eight mammalian enzymes), and can be erased by KDM4A,KDM4B, KDM4C or NO66[32]. Moving from 5' to 3' ends of genes, several groups have observed progressive methylation of H3K36 (me1 to me3), and multiple lines of evidence suggest it is important for maintaining transcriptional elongation and preventing spurious initiation in gene bodies[34,35].

Among inactive marks, methylation of H3K9 and H3K27 are the most well-studied. H3K9me3 is mostly known for its role in silencing heterochromatin, and can be written by SUV39H1/2, SETDB1, or PRDM2, and erased by several KDM family members including KDM3B, KDM4A, KDM4B, KDM4C, and KDM4D in humans[32]. In particular, H3K9me3 is found at high levels in constitutive heterochromatin, regions of the genome enriched in repeats, including centromeric and telomeric satellites, retrotransposons and endogenous retroviruses, that must be kept silent and compact in all cell types to protect genomic integrity. However, H3K9me3 is also important in silencing facultative heterochromatin, genomic regions containing promoters and enhancers with cell-type restricted activity. Indeed, during differentiation H3K9me3 spreads to form long heterochromatin regions up to megabases in scale[36]. H3K9me3 is largely mutually exclusive with H3K27me3, the mark of the Polycomb complex.

Polycomb group proteins, so named because *Drosophila* mutants of *polycomb* develop multiple sex combs, form essential silencing complexes conserved throughout eukaryotes. In mammals, many distinct Polycomb complexes exist, but can be broadly divided into two major familes: Polycomb Repressive Complex 1 & 2 (PRC1/2). PRC1 has as a core component RING1B, an E3-ubiquitin ligase that is responsible for H2AK119Ub1, which is recognized by

JARID2, an accessory to the PRC2 complex containing EZH1/2 that catalyze di- and trimethylation of H3K27, although Polycomb recruitment can work through several distinct mechanisms. In *Drosophila*, polycomb is targeted by so-called Polycomb Response Elements (PREs), but in mammals appears to be recruited primarily by CGI[37]. At CGI, H3K27me3 is repressive to transcriptional initiation and temporally predisposes to eventual DNA methylation, although H3K27me3 and DNA methylation rarely co-occur suggesting they are distinct silencing mechanisms (with DNA methylation perhaps more permanent). In stem cells, H3K4me3 and H3K27me3 often occur together at bivalent domains that may undergo subsequent activation or sustained repression.

Modern techniques for interrogating histone modifications and variants rely on specific antibodies to enrich chromatin containing the mark of interest. The DNA is then isolated and subject to high throughput sequencing in a technique known as ChIP-Seq, although earlier iterations relied on PCR of specific loci or use of a microarray (ChIP-chip). Several multinational consortiums have performed ChIP-Seq for the most important marks in a variety of cancer and normal cell types, leading to a wealth of publicly available data useful for correlations in genome-wide experiments.

Crosstalk between DNA Methylation and Histone Modifications

DNA methylation is intrinsically related to histone modifications, with many epigenetic writers and readers directly recruited or repelled by DNA methylation, and many chromatin marks and modifiers in turn attracting or disrupting DNMT. As such, CGI exhibit striking levels of certain active marks like H3K9/27Ac and H3K4me3 that are associated with transcriptional initiation. For example, Cfp1 (CXXC finger protein 1), and many MLL (Mixed-lineage leukemia) family members, H3K4me3 methylases, contain CXXC domains that explicitly target them to unmethylated CpG sites[38]. In turn H3K4me acts to prevent the activating effects unmodified H3K4 has on DNMT3A[39], an effect that can be mediated by DNMT3L, which senses unmethylated H3 tails via its ADD domain to recruit activated DNMT3A and *de novo* methylation[40].

Just as CGI recruit H3K4me, they also have mechanisms to remove H3K36me3, which represses initiation. The H3K36me3 demethylases JMDH1A & JMDH1B also possess CXXC domains that target to CGI. However, in gene bodies DNMT3A is recruited to H3K36me3 via it's PWWP domain, facilitating DNA methylation[35].

In pericentric heterochromatin, heterochromatin protein 1 (HP1) specifically binds to H3K9me3 and recruits DNMT3B[41], critical to the silencing of major satellites. In a feedback loop, MBD1, which contains a methyl-binding domain (MBD) recognizing methylated CpG sites, can form a complex with HP1 and SUV39H1/H2, or with SETDB1, to target them to methylated regions. Furthermore, SETDB1 itself has two putative MBD domains that may also target it to methylated sites[40].

Of course, histone modifications also exhibit substantial regulation of each other. For example PHF8 recognizes H3K4me2/3, but catalyzes demethylation of H3K9me2, apparently playing a role in preventing repressive H3K9me in promoters. Similarly, JHDM1D also binds H3K4me2/3, but demethylates H3K27me3, preventing polycomb silencing. Likewise, an H3K4me3 methylase MLL2-MLL3 complex associates with another H3K27me3 demethylase known as UTX. In contrast, both polycomb group proteins and H3K9 methyltransferase complexes contain H3K4me2/3 demethylases, and often HDACs as well. Histone marks can also reinforce each other: monoubiquitination of H2B is required for both H3K4me and H3K79me3, a mark associated with transcriptional elongation[42].

Transcription and Pol II Pausing

Arguably, the primary role of the epigenetic code is dictating transcription, and faithfully preserving encoded expression programs across mitotic generations. Eukaryotes possess three primary RNA polymerase complexes: Pol I which transcribes rRNA, Pol II, responsible for mRNA, long noncoding RNA (lncRNA), eRNA, and upstream antisense RNA (uaRNA), and Pol III which produces tRNA and some other small RNAs. Ultimately, the spatial and temporal modulation of Pol II by the epigenetic code is responsible for the vast majority of phenotypic differences between cell types and how cells respond to environmental stimuli. To initiate transcription at its targets, Pol II, a 12-subunit enzyme, unites with the general transcription factors (GTFs: TFIIB, TFIID, TFIIE, TFIIF, and TFHIIH) in a pre-initiation complex (PIC) which melts DNA to form a 'transcription bubble', commence transcription, and secure release from many of the GTFs in a process known as 'promoter escape'[43].

At this point transcription continues for 20-60bp until stopping in a process known as promoter-proximal Pol II pausing. While the central dogma was established in the late 1950s, it was not clear that significant post-initiation regulation of transcription occurred until the late 1970s with the Lis laboratory's studies of the *Drosophila* heat shock loci. At these genes, they discovered that Pol II accumulates just downstream of the transcription start site (TSS) along with nascent RNAs. While there were hints of widespread Pol II pausing over the next decades, it was not until the advent of ChIP-Seq (see below) that it became clear that Pol II exhibits much higher density near promoters than throughout gene bodies, indicating that Pol II pausing is a near-universal mechanism of transcriptional regulation. GroSeq and ProSeq have recently provided an even clearer portrait of Pol II pausing, and we now know that CGI in particular are characterized by high initiation levels, although pausing can largely prevent progression of Pol II into the gene body.

DNA methylation generally prevents initiation at promoters. CGI in general are only capable of initiation when unmethylated. Transcription in non-CGI promoters, which display moderate to heavy methylation and much less initiation overall than unmethylated CGI, also negatively correlates with methylation.

Pol II pausing is enforced by two factors: NELF (negative elongation factor) and DSIF (doxorubicin-sensitivity inducing factor). The largest subunit of Pol II possesses a C-terminal domain (CTD) consisting of ~52 heptad repeats, and paused Pol II exhibits phosphorylation of Ser-5 of this repeat. To be released from pausing, Pol II must recruit active P-TEFb kinase (positive transcription elongation factor, composed of CDK9 and cyclin T1, T2 or K). P-TEFb can exist either inactively in a complex with HEXIM1 or HEXIM2, mePCE, LARP7, and small nuclear ribonucleoprotein (snRNP) 7SK or separately from these factors in an active complex with certain TFs. Recruitment and activation of P-TEFb depends on many factors, most prominently binding of MYC or looping with enhancers, especially in the context of interaction with BRD4, a protein containing several bromodomains that physically bridges histone acetylation at enhancers and promoters as well as competing with the repressive HEXIM1/2 complex for P-TEFb binding, leading to P-TEFb activation. P-TEFb can then phosphorylate NELF/DSIF as well as Ser2 of the Pol II CTD, permitting dissociation of NELF, while DSIF continues to travel with released Pol II. Pol II pausing is thought to play important roles in gene regulation, providing an additional means of modulating mRNA levels between cell types or conditions, maintaining open chromatin at promoters, synchronizing or priming activation in response to developmental or environment stimuli, and permitting each nascent transcript time to acquire the 5' 7-methylguanylate cap necessary for RNA stability.

Pol II can encounter numerous other barriers to elongation as it progresses through gene bodies. Sequence features like exons, long terminal repeats, and regions of high GC or

CpG content can impede Pol II, while H3K79me2 and H4K20me1 appear to accelerate elongation rate[44]. To trigger transcriptional termination at 3' gene ends, poly A sites, as well as perhaps R-loops (DNA:RNA hetroduplexes that form co-transcriptionally especially in regions of high GC-skew) and H3K9me2/3 help to dramatically slow Pol II for polyadenylation and processing.

Highly expressed genes tend to have methylated gene bodies, and in many organisms levels of DNA methylation in the gene body positively correlate with expression. Intragenic DNA methylation is thought to silence spurious promoters and repetitive elements or retroviruses present in genes. Exons also tend to be more methylated than introns, and DNA methylation may play a role in splicing regulation. Indeed, treatment with the DNMT inhibitor decitabine results in changes to exon inclusion in cells[45]. As well, many genes rely on DNA methylation for full expression: transient decitabine treatment lowers the expression of many genes that cannot regain their endogenous expression following drug withdrawal without DNMT3A[46].

Transcriptional initiation by Pol II is not limited to protein-coding mRNAs. Recent work has demonstrated most initiation sites exist in bidirectional pairs. At genes, mRNAs are typically paired with an upstream antisense transcripts, the roles of which are unclear but are thought to maintain an open transcriptional state at promoters. Similarly, at enhancers, Pol II initiated birectionally, but both eRNA transcripts are typically unstable[47].

Enhancers

Enhancers are genomic regions that act at a distance to promote gene transcription independent of position or orientation. Enhancers are similar to CGI in that their levels of DNA methylation are inversely correlated with their activity and eRNA levels[48–51]. Indeed, recent work has demonstrated that enhancers and promoters possess a unified chromatin architecture, consisting of H3K4me, H3K9/27Ac, DNAse hypersensitivity, paired bidirectional transcriptional initiation, and transcription factor binding [52–54] . Measures of transcript stability most reliably distinguish the two, with promoters displaying one or two stable transcripts, and enhancers strictly unstable (eRNA) transcripts [54].

The first enhancer identified was a 72-bp region of the simian virus SV40 that when cloned behind a reporter massively enhanced transcription, and subsequent studies in the 1980s focused on the immunoglobin loci revealed extensive regulation of gene expression by endogenous enhancers in mammals[54]. Enhancers form physical loops with promoters, and are thought to work by bringing the TFs, chromatin modifiers and remodelers, and transcriptional machinery that they recruit to promoters. Indeed many TFs including cancer-relevant estrogen receptor (ER) and FOXA1 appear to preferentially bind enhancers vs promoters, suggesting that enhancers are critical to connecting multiple signaling events to gene activation [55].

Genome-wide methods to identify enhancers have typically focused on detecting TF binding, DNAse hypersensitivity, histone acetylation or HAT presence, and H3K4me1; in particular overlapping ChIP-Seq peaks of H3K27Ac and H3K4me1 are typically thought to represent enhancer regions [56]. Such methods predict hundreds of thousands of enhancers in the human genome, but the functional relevance of the vast majority of putative enhancers has not been tested experimentally. Regardless, the sheer number of enhancers relative to the ~20 thousand promoters in the genome suggest that enhancers play a universally critical role in programming transcription.

As stated above, most if not all enhancers exhibit transcriptional initiation. In contrast to stable mRNAs, but like unstable uaRNAs, enhancers have enrichment in polyadenylation sites, but a lack of stabilizing U1 snRNP motifs. Stabilizing features lead to protection from degradation by the RNA exosome, while destabilizing features lead to rapid degradation of the nascent RNA. Thus, stable transcripts can be readily detected at steady-state levels, marking mRNA and lncRNA, among others, while only methods capturing nascent transcription can capture unstable transcripts like eRNA and uaRNA.

Like mRNAs, eRNAs appear dependent on release from Pol II pausing and exhibit substantial P-TEFb recruitment and often form transcripts several hundred base pairs in length. The biological function of these eRNAs is not clear. Some have speculated that eRNAs may simply represent transcriptional noise via sampling of open regions of the genome by Pol II and the proximity enhancers exhibit to promoter-based transcriptional machinery. Transcription itself may also play a role in maintaining an open chromatin structure for TF binding at enhancers. However, knockdown of specific eRNAs has been shown to reduce target gene expression as well as physical looping of the enhancer to the promoter. A role for eRNA in releasing Pol II pausing has also been demonstrated as eRNA can compete with NELF-E for binding to the nascent RNA [57].

Transcription genome-wide is most commonly assayed by simple sequencing of mRNAs, with variations on the size and nature of RNAs selected, known as RNA-Seq. While RNA-Seq provides an excellent picture of steady-state mRNA levels, it is poorly equipped to assay nascent transcription, and misses highly paused or unstable transcripts. However, global run-on sequencing (Gro-Seq) [58] and the related precision run-on sequencing (Pro-Seq)[59] have been developed to study actual transcriptional dynamics in cells. In these techniques, nuclei are isolated and new Pol II initiation is halted by sarkosyl treatment. Transcription is then allowed to 'run-on' in the presence of labeled nucleotides, bromo-uridine for GroSeq which allows run-on of several hundred bases, and biotinylated ATP, CTP, GTP, and TTP in ProSeq which halt transcription after incorporation giving base-resolution of engaged Pol II genome-wide. Thus ProSeq in particular allows visualization of Pol II pausing dynamics, uaRNA, eRNA, and other transcripts not present at substantial steady-state levels.

Interestingly, half of CGI do not in fact overlap known TSSs and thousands lie far from any known transcript. While many of these 'orphan' CGI are likely promoters of unannotated transcripts, their prevalence suggests there may be roles for CGI other than as strict promoters, and several groups have noted enhancer-like chromatin features at orphan CGI.

Epigenetic Aberrations in Cancer

While it has been known for decades that genetic mutation is a primary cause of cancer, we now know that epigenetic changes are critical in tumor initiation and progression. Pervasive DNA hypomethylation was the first epigenetic phenotype characterized in cancer cells in 1984 [60], a finding since confirmed in nearly ever tumor type with functional consequences for genomic instability and overexpression of key oncogenes [61]. In spite of global hypomethylation, DNA hypermethylation events were soon also linked to cancer with the observation that the DNA repair enzyme MLH1 was silenced by promoter methylation in colon cancers [62]. Indeed, promoter hypermethylation has been demonstrated at dozens of tumor-suppressor genes including *TMS1/PYCARD, RB, CDKN2A, VHL*, and *CDH1*, and 5-10% of all promoter CGI exhibit hypermethylation in tumors. Outside of strict promoter silencing, changes in methylation in the regions surrounding CGI ('shores') are also frequent in cancer. While the role of shore methylation is less defined, these changes may modulate TF binding to selectively alter transcription in cancer cells [63,64]. The bodies of over-expressed genes can also gain DNA methylation including *MGMT* which is responsible for temozolomide resistance in gliomas [65]. Enhancers have been suggested to exhibit more cancer-related methylation changes than other compartments, and in melanoma, enhancer methylation appears to play a role in adapting the primary tumor to metastasis [66,67,50]. Cancer cells also tend to exhibit global loss of hydroxymethylation [68], which may also have important implications for enhancer function.

Redistribution of histone marks in cancer is as pervasive as transcriptional changes, and many tumors exhibit global gains or losses in certain modifications. Often such gain or loss is linked to genetic mutation of epigenetic regulators, which occurs in every family of chromatin modifying enzyme and epigenetic mark reader across tumor types. Indeed, a recent analysis suggests that at least 50% of all human cancers harbor a mutation in an epigenetic regulator [61]. For example, hematological malignancies like acute myeloid leukemia (AML), diffuse large B-cell lymphoma (DLBCL), and acute lymphoblastic leukemia (ALL) often exhibit mutation in the HATs KAT3A (CBP) and KAT3B (p300), and ALL often also has mutations in the acetylation reader BRD1. AML also frequently possesses mutated histone methyltransferases like MLL1, KMT3B/NSD1, and NSD3, while NSD2 mutations characterize multiple myeloma. Breast cancers often contain mutated KMT2C/MLL3, the methylation reader ING1, the acetylation reader PBRM1, and the HAT KAT3B. Even DNA methylation can be genetically disrupted via mutation of DNMT3A, TET1, and TET2 in AML and other tumors [61]. The genes encoding histones themselves can also be mutated: H3.3K27M mutations act as a dominant negative to inhibit polycomb activity genome-wide [69] ,and H3.3K36M mutations drive sarcomas and chrondroblastomas [70] by preventing H3K36me3.

Epigenetic Therapy

Unlike the genetic mutations that contribute to cancer, the epigenetic alterations are reversible and thus represent provocative drug targets. Over 40 years ago, the first epigenetic therapy to be shown to treat cancer was the DNMT inhibitor 5'-azacitidine (AZA), a cytidine analog that is incorporated into DNA where it traps DNMT when it undertakes a futile effort to methylate AZA's nitrogen substitution at carbon-5. This mechanism was unclear at the time and AZA as developed as a simple nucleoside analog. The DNA analog of AZA, 5-aza-2′-deoxycytidine (DAC or decitabine), works through a similar mechanisms, but is only incorporated into DNA while AZA is integrated into both DNA and RNA. Both are FDA approved for the treatment of myelodysplastic syndrome and DAC is approved for acute refractory AML [61].

While relatively ineffective as single agents against solid tumors in patients, DNMT inhibitors do decrease the proliferation rate and invasive potential of breast, lung and colon cancer cells in culture and mouse models. Intriguingly, these effects appear to result from durable reprogamming of the cancer epigenome, as these effects last for months following drug withdrawal. While the activity of such drugs has been attributed to the activation of tumor-suppressors silenced by promoter methylation, it is also clear the gene body methylation at oncogenes including *MYC* is a therapeutic target [46]. Recently, it was shown that the activation of certain endogenous retroviruses (ERVs) by AZA/DAC-induced demethylation resulted in ERV dsRNA accumulation in the cytosol and the triggering of a signaling cascade mediated by the dsRNA sensor MAVS to dampen growth rate [71].

HDAC inhibitors, originally found in screens for compounds that could differentiate leukemia cells [61], have also found use in the treatment of tumors, as well as psychiatric illnesses. For example, vorinostat, romidepsin, and belinostat are currently used to treat T-cell lymphomas. Additional trials have demonstrated synergy of HDAC and DNMT inhibition which augments the gene activation potential of each drug. Such combination therapy may even sensitize cancer cells to traditional cytotoxic drugs and immunotherapies.

EZH2 (writer of H3K27me3) inhibitors are currently in clinical trials for several cancers that over-express the gene or possess activating mutations, as are DOT1L (writer of H3K79me3) inhibitors in leukemias with MLL-rearrangements. LSD1 (eraser of H3K4me and H3K9me) inhibitors are also demonstrating efficacy against AML and small-cell lung cancer. Overall more than 30 epigenetic drugs are currently in various phases of clinical trials and epigenetic therapy is a promising field in cancer treatment[61] .

Objectives

In this thesis, I address a number of questions critical to understanding both the mechanisms of epigenetic therapy and the intrinsic biology of DNA methylation, specifically its role in regulating transcription, and how CGI, unique in their intense hypomethylation, in particular represent critical functional elements in the genome. First, I focus on understanding how decitabine durably reprograms the cancer epigenome, addressing the genomic and epigenomic features that determine why certain CpGs regain lost methylation immediately, while others never do. This question is vital to understanding the biology of relapse following decitabine treatment, but also represents an experimental approach to assaying inherent susceptibility to *de novo* methylation. Next, I address the relationship between nascent transcription and CGI promoters. Most CGI exhibit substantial Pol II initiation levels, yet the clear variation in actual mRNA production among associated genes suggests that CGI may have unique barriers to transcriptional elongation. Third, I address the longstanding question of orphan CGI function. I undertake an extensive analysis of such CGI, integrating an armada of genetic and epigenetic evidence to test the hypothesis that they function as enhancers. Finally, I utilize ProSeq to specifically answer how DNA methylation regulates nascent transcription genome-wide and document how epigenetic therapy returns the cancer transcriptional profile to a more normal state.

Bibliography

1. Waddington, C. H. The epigenotype. *Int. J. Epidemiol.* **41,** 10–13 (2012).
2. Van Speybroeck, L., De Waele, D. & Van De Vijver, G. Theories in early embryology: close connections between epigenesis, preformationism, and self-organization. *Ann. N. Y. Acad. Sci.* **981,** 7–49 (2002).
3. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* **79,** 137 (1944).
4. Lyon, M. F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *nature* **190,** 372–373 (1961).
5. Ohno, S., Kaplan, W. D. & Kinosita, R. Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus. *Exp. Cell Res.* **18,** 415–418 (1959).
6. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Genome Res.* **14,** 9–25 (1975).
7. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *COLD SPRING Harb. Monogr. Ser.* **32,** 639–645 (1996).
8. Lee, J. T. & Bartolomei, M. S. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* **152,** 1308–1323 (2013).
9. Feng, S., Jacobsen, S. E. & Reik, W. Epigenetic reprogramming in plant and animal development. *Science* **330,** 622–627 (2010).
10. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark

and transcription initiation at most promoters in human cells. *Cell* **130,** 77–88 (2007).

11. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449,** 248–251 (2007).
12. Ooi, S. K. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448,** 714–717 (2007).
13. Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet* **7,** e1002389 (2011).
14. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99,** 247–257 (1999).
15. Jones, P. A. & Liang, G. Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.* **10,** 805–811 (2009).
16. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy) methylcytosine and its oxidized derivatives. *Cell* **152,** 1146–1159 (2013).
17. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40,** 91–99 (1985).
18. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
19. Brandels, M., Frank, D., Mendelsohnf, M. & Nemesf, A. Sp1, elements protect a CpG. *Nature* **371,** 29 (1994).
20. Macleod, D., Charlton, J., Mullins, J. & Bird, A. P. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8,** 2282–2292 (1994).
21. Kellner, W. A., Bell, J. S. & Vertino, P. M. GC skew defines distinct RNA polymerase pause sites in CpG island promoters. *Genome Res.* **25,** 1600–1609 (2015).
22. Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a

connection with DNA methylation. *Genes Dev.* **13,** 1924–1935 (1999).

23. Baylln, S. B., Herman, J. G., Graff, J. R., Vertino, P. M. & Issa, J.-P. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv. Cancer Res.* **72,** 141–196 (1997).
24. Issa, J. P., Vertino, P. M., Boehm, C. D., Newsham, I. F. & Baylin, S. B. Switch from monoallelic to biallelic human IGF2 promoter methylation during aging and carcinogenesis. *Proc. Natl. Acad. Sci.* **93,** 11757–11762 (1996).
25. Conway, K. E. *et al.* TMS1, a novel proapoptotic caspase recruitment domain protein, is a target of methylation-induced gene silencing in human breast cancers. *Cancer Res.* **60,** 6236–6242 (2000).
26. Grady, W. M. *et al.* Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nat. Genet.* **26,** 16–17 (2000).
27. Esteller, M., Hamilton, S. R., Burger, P. C., Baylin, S. B. & Herman, J. G. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer Res.* **59,** 793–797 (1999).
28. Kane, M. F. *et al.* Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Res.* **57,** 808–811 (1997).
29. Medvedeva, Y. A. *et al.* Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* **15,** 119 (2014).
30. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17,** 208 (2016).
31. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci.* **51,** 786–794 (1964).
32. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–357 (2012).

33. Verdin, E. & Ott, M. 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. *Nat. Rev. Mol. Cell Biol.* **16,** 258–264 (2015).

34. Venkatesh, S., Li, H., Gogol, M. M. & Workman, J. L. Selective suppression of antisense transcription by Set2-mediated H3K36 methylation. *Nat. Commun.* **7,** (2016).

35. Wagner, E. J. & Carpenter, P. B. Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.* **13,** 115–126 (2012).

36. Becker, J. S., Nicetto, D. & Zaret, K. S. H3K9me3-dependent heterochromatin: barrier to cell fate changes. *Trends Genet.* **32,** 29–41 (2016).

37. Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* **1,** e1500737 (2015).

38. Lee, J.-H. & Skalnik, D. G. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.* **280,** 41725–41731 (2005).

39. Guo, X. *et al.* Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature* **517,** 640–644 (2015).

40. Cheng, X. Structural and functional coordination of DNA and histone methylation. *Cold Spring Harb. Perspect. Biol.* **6,** a018747 (2014).

41. Lehnertz, B. *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr. Biol.* **13,** 1192–1200 (2003).

42. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–357 (2012).

43. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 129–143 (2015).

44. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features

and epigenetic modifications. *Genome Res.* **24,** 896–905 (2014).

45. Ding, X.-L., Yang, X., Liang, G. & Wang, K. Isoform switching and exon skipping induced by the DNA methylation inhibitor 5-Aza-2′-deoxycytidine. *Sci. Rep.* **6,** (2016).
46. Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26,** 577–590 (2014).
47. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).
48. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).
49. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–220 (2013).
50. Aran, D. & Hellman, A. DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* **154,** 11–13 (2013).
51. Barwick, B. G., Scharer, C. D., Bally, A. P. & Boss, J. M. Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation. *Nat. Immunol.* **17,** 1216–1225 (2016).
52. Chen, R. A.-J. *et al.* The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. *Genome Res.* **23,** 1339–1347 (2013).
53. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155,** 990–996 (2013).
54. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).
55. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12,** 283–293 (2011).

56. Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: Roadmap for regulation. *Nature* **518,** 314–316 (2015).

57. Liu, W. *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155,** 1581–1595 (2013).

58. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12,** 433–438 (2015).

59. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11,** 1455–1476 (2016).

60. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4,** 143–153 (2004).

61. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150,** 12–27 (2012).

62. Herman, J. G. *et al.* Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc. Natl. Acad. Sci.* **95,** 6870–6875 (1998).

63. Rao, X. *et al.* CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene* **32,** 4519–4528 (2013).

64. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41,** 178–186 (2009).

65. Moen, E. L., Stark, A. L., Zhang, W., Dolan, M. E. & Godley, L. A. The role of gene body cytosine modifications in MGMT expression and sensitivity to temozolomide. *Mol. Cancer Ther.* **13,** 1334–1344 (2014).

66. Bell, R. E. *et al.* Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.* gr. 197194.115 (2016).

67. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14,** 1 (2013).

68. Kroeze, L. I., van der Reijden, B. A. & Jansen, J. H. 5-Hydroxymethylcytosine: An epigenetic mark frequently deregulated in cancer. *Biochim. Biophys. Acta BBA-Rev. Cancer* **1855,** 144–154 (2015).

69. Lewis, P. W. *et al.* Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. *Science* **340,** 857–861 (2013).

70. Lu, C. *et al.* Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science* **352,** 844–849 (2016).

71. Chiappinelli, K. B. *et al.* Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162,** 974–986 (2015).

Chapter II: GC skew defines distinct RNA polymerase pause sites in CpG island promoters

Wendy A. Kellner*, Joshua S.K. Bell*, Paula M. Vertino

* These authors contributed equally to this work

Published in Genome Research 2015 Nov; 25(11), and is included here as published.

Contributions: WK, JB, and PV wrote the manuscript. WK and JB performed the data analysis, and WK performed the 3C experiments.

Abstract

CpG islands (CGI) are associated with over half of human gene promoters and characterized by a unique chromatin environment and high levels of bidirectional transcriptional activity relative to surrounding genomic regions, suggesting that RNA polymerase (Pol II) progression past the CGI boundaries is restricted. Here we describe a novel transcriptional regulatory step wherein Pol II encounters an additional barrier to elongation distinct from the promoter-proximal pause and occurring at the downstream boundary of the CGI domain. For most CGI associated promoters, Pol II exhibits a dominant pause at either the promoter-proximal or this distal site that correlates, both in position and in intensity, with local regions of high GC skew, a sequence feature known to form unique secondary structures. Upon signal–induced gene activation, long-range enhancer contacts at the dominant pause site are selectively enhanced, suggesting a new role for enhancers at the downstream pause. These data point to an additional level of control over transcriptional output at a subset of CGI associated genes that is linked to DNA sequence and the integrity of the CGI domain.

Introduction

Roughly 60% of human promoters are associated with a CpG island (CGI), most of which lack DNA methylation and maintain a chromatin structure that is permissive to transcription; the acquisition of DNA methylation at a small percentage of these promoters during development or disease is associated with stable gene silencing (Deaton and Bird 2011; Jones 2012). Histone modifying enzymes contain embedded or associated reader domains capable of recognizing methylated or unmethylated CpGs, allowing for crosstalk between DNA methylation state and local chromatin structure (Hashimoto et al. 2010). For example, CGI are maintained in a transcriptionally permissive state in part through the recognition of unmethylated DNA by a component of the H3K4 methyltransferase complex and the inability of de novo DNA methyltransferases to act on H3K4 modified chromatin (Jia et al. 2007; Thomson et al. 2010). As a result, there is an inverse relationship between DNA methylation and H3K4 methylation, with unmethylated CGI domains uniquely marked by H3K4me3 genome-wide. DNA sequence features have also been reported to promote or to prevent DNA methylation at CGI (Feltus et al. 2003; Bock et al. 2006; Ginno et al. 2012). How chromatin structure and DNA sequence converge to regulate transcription initiation and elongation at CGI is not well understood.

Genome-wide studies of RNA polymerase (Pol II) occupancy and nascent transcription have demonstrated that a significant component of transcriptional regulation occurs at post-initiation steps in the transcription cycle. Promoter-proximal pausing has emerged as an important point of post-initiation transcriptional regulation that is conserved across metazoans (Adelman and Lis 2012; Kwak and Lis 2013). After transcribing ~50bp, initiated Pol II pauses awaiting additional signals for controlled release into productive elongation. This allows for rapid and/or synchronous gene activation in response to a wide variety of environmental or developmental cues. In most cases, elongation past this point requires the recruitment of positive transcription elongation factor B (P-TEFb), which phosphorylates the C-terminal domain of Pol II as well as the negative regulatory factors NELF and DRB sensitivity-inducing factor (DSIF), promoting their dissociation/inactivation and the release of Pol II into active elongation (Gilchrist et al. 2010). While transient pausing is thought to be a feature of most active transcription, the degree to which this step becomes rate-limiting varies across genes, and is subject to context-dependent and locus-specific modulation, presumably by factors affecting the local recruitment and/or activity of P-TEFb. Central among these is bromodomain-containing protein 4 (BRD4) which directs P-TEFb to acetylated nucleosomes while also antagonizing its sequestration by the HEXIM1 complex (Jang et al. 2005; Yang et al. 2005; Liu et al. 2014). Recent studies suggest that distal enhancer interactions play a key role in mediating these events. Enhancers are *cis*-acting regulatory elements that control transcription from a distance through the formation of contacts between the enhancer-bound transcription factors and promoter-bound Pol II and the looping out of intervening chromatin (Kagey et al. 2010b). BRD4 and P-TEFb have been shown to co-localize with elongating (serine-2-phosphorylated) Pol II at both the promoters and enhancers of active genes, and inhibition of either suppresses elongation not only of promoter-derived mRNAs but also that of non-coding ‘eRNAs’ arising at distal enhancers (Zhang et al. 2012; Anand et al. 2013; Liu et al. 2013; Loven et al. 2013; Kanno et al. 2014). In addition, long-range chromatin looping interactions have been shown to correlate with paused Pol II during *Drosophila* development (Ghavi-Helm et al. 2014). The relationship between chromatin looping interactions, enhancer activity and Pol II pausing dynamics are incompletely understood.

In this study, we investigate the relationship between DNA sequence features, chromatin structure and RNA Pol II pausing dynamics in the regulation of transcription at CGI promoters. We identify a novel Pol II pause point distinct from the promoter-proximal pause defined by local DNA sequence features that is coincident with the downstream edge of CGI domain and serves as the predominant barrier to elongation at a significant fraction of CGI associated genes. We further find that contacts between the distal enhancers and the predominant Pol II pause points are selectively enhanced upon stimulation, implicating a new role for enhancers at this distal site.

Results

Genome-wide mapping of nascent transcription has shown that unmethylated CpG island promoters support high levels of bidirectional (divergent) transcription, but productive elongation in only one direction, and that the levels of activity in the promoter region are poorly correlated with the levels of steady state transcript (Core et al. 2008). This implies that there may be additional levels of transcriptional regulation at the CGI boundaries. To address this question, we performed an integrated analysis of global run-on sequencing (GRO-seq) (Hah et al. 2011), methylated DNA immunoprecipitation sequencing (MeDIP-seq) (Ruike et al. 2010), Pol II ChIP-seq (Lee et al. 2012), and ChIP-seq of histone modifications from human MCF7 cells at TSSs falling within a CGI or not. As expected, CGI-associated promoters exhibit high levels of divergent transcription and the absence of DNA methylation that distinguishes these from non-CGI promoters (**Supplemental Figure 1**). Relative to non-CGI promoters, CGI promoters have significantly more engaged Pol II and are enriched in histone modifications associated with ongoing transcription including H3K4me3 and acetylated H3 (**Supplemental Figure 1C, D**). Histone modifications and transcriptional activity are distributed on both sides of the TSS, indicating that these marks are associated with both sense and divergent transcription at CGI promoters. This analysis also highlights that the majority of paused genes and ChIP-seq signal from chromatin modifications associated with transcriptional activation derives from CGI promoters.

To examine the relationship between chromatin structure and engaged Pol II specifically at CGI, promoter-associated CGI were sorted by CGI size. Transcriptional activity, indicated by GRO-seq tag density, was confined to the unmethylated CGI domain and corresponded with Pol II enrichment in this region (**Figure 1A**). Histone modifications known to be associated with active transcription, such as H3K4me3 and H3K9ac, were similarly concentrated within the unmethylated CGI domain (**Supplemental Figure 2**). Interestingly, there was significant GRO-seq enrichment at the edges of the CGI domain. Sense strand transcripts are enriched at the downstream edge of CGI and antisense strand transcripts are enriched at the upstream edge, suggesting that Pol II may pause in these regions. These data confirm that divergent transcription is a common feature of most if not all active CGI-associated promoters (Core et al. 2008), and underscores the relationship between divergent transcription and surrounding chromatin at CGIs, ie. divergent transcription appears largely confined to the CGI domain.

We next sought to determine the relationship between divergent transcription, promoter-proximal pausing and DNA methylation at CGI promoters, while taking into consideration the relative position of the TSS within the CGI domain. Genes were oriented to the direction of transcription and normalized to the CGI length such that the distance to the upstream and downstream CGI edges were independently scaled relative to the TSS. Meta-analysis across 16,657 CGI-associated promoters revealed an enrichment of GRO-seq sense strand tags at ~50bp downstream of the TSS, representing the well-characterized promoter-proximal pause (**Figure 1B**). Interestingly, we observed a second distinct accumulation of nascent transcripts (sense strand) at the downstream edge of the CGI domain (second blue peak, **Figure 1B**). Notably, divergent transcription was similarly confined to the region upstream of the TSS to the 5’ CGI edge. Taken together, these data indicate that Pol II encounters a barrier to transcription past the downstream edge of the CGI domain that is distinct from the promoter proximal-pause, and further, that this position represents a major regulatory step for continued elongation at a significant number of genes. We refer to this as the ‘distal’ pause.

The above observations led us to inquire whether both pausing events were characteristic of most genes; i.e. whether most genes are regulated by two sequential pauses, or whether the average profile might represent distinct groups of genes with different pausing characteristics. We therefore parsed genes by the relative ratio of the proximal versus distal pausing indices, defined as the GRO-seq tag density of the 100 bp encompassing the proximal or distal peak relative to that of the gene body downstream of the CGI. This analysis indicated that CGI-associated genes can be grouped into two classes, those predominantly regulated at the proximal site or those regulated at the distal site (**Figure 1C**). Approximately 35% of CGI associated genes showed predominant pausing at the distal position (Class I, distally-paused), whereas 47% showed a more prominent proximal pause (Class II, proximally-paused). A third class (18%) lacked significant GRO-seq tags indicating that they are silent in this cell type. Limiting the analysis to only those CGI that contain a single annotated TSS had no impact on this pattern and the same relationships were observed (**Supplemental Figure 2B**). For many genes there is a dominance of one pause point over the other, although both pauses are still observed in the other location (for example *HSPA4* and *ESR2)*, while other genes, like *MYC* and *FOS* show a clear preference for either the distal or proximal pause, respectively (**Figure 1D**). Thus, it would appear that Pol II encounters a proximal and distal pause at most CGI genes, but the degree to which each is rate-limiting varies, with one or the other dominating.

Because the distal pause coincides with the edge of the CGI, we reasoned that DNA methylation and/or histone modifications might be candidates for regulating pausing at this position. However, analysis of histone modifications correlated with promoters and/or transcription (H3K4me3, H3K9ac, H3K14ac), or shown in Figure S2 to be associated with the CGI edge (H3K4me2, H3K9me1, DNA methylation) showed no difference in enrichment patterns between the pausing classes (**Supplemental Figure 3A**). Several protein factors implicated in the regulation of promoter-proximal pausing, such as CCNT2 and CDK9 (components of the P-TEFb complex), MYC (Rahl et al. 2010), and BRD4 (Delmore et al. 2011), were similarly analyzed. Although there was clear enrichment of these factors at the proximal pause point just downstream of the TSS as expected, this pattern was common across all active CGI, and there was no difference in profiles between the pausing classes (**Supplemental Figure 3B**). Whereas steady state gene expression, gene size, and fraction of the gene encompassed within the CGI did not significantly differ between CGI classes, distally paused genes tended to be associated with larger CGI and downstream distances from the TSS to 3'CGI edge, and a higher CpG density than proximally paused genes, although the differences were relatively modest (median CGI size = 871bp vs. 944bp vs. 1018bp, proximal, distal, silent, respectively;median distance TSS to 3' CGI edge = 498bp vs. 531bp vs. 570bp, proximal, distal, silent, respectively) (**Supplemental Figure 4**). Interestingly, the silent class shared some features with the distal class (slightly longer CGI and downstream distance) but represented a class of CGI with lower CpG density (**Supplemental Figure 4**). Gene ontology and gene set enrichment analyses indicated that the genes in the proximal and distal classes are similar in terms of function. While both classes were enriched in housekeeping functions, there was a tendency for the proximal class to be enriched in transcriptional and RNA processing activities , while the distal class tended towards cell cycle and metabolic processes (**Supplemental Data**). The most striking finding was in the silent class, which is strongly enriched in genes subject to Polycomb-mediated repression in stem cells and other tissues.

We next asked whether other cell types with different expression patterns have the same pausing patterns. The profile of ongoing transcription (GRO-seq) from normal breast epithelia (MCF10A) (Kim et al. 2013) and fetal lung fibroblast (IMR90) (Core et al. 2008), and ChIP-seq for serine-5 phosphorylated (initiated) Pol II from CD4+ T-cells (Zhang et al. 2012) was examined at CGI-associated promoters using the same sort order (distance from the TSS to the 3’ CGI edge) and class distinctions derived for MCF-7 cells in Fig 1C. Interestingly, the position of the dominant pause for individual genes was consistent across multiple cell types (**Figure 2**). These data suggest that the predominant pausing class is intrinsically determined, and independent of cell-type specific expression patterns.

Core promoter sequence elements have been implicated in the regulation of promoter-proximal pausing (Kwak et al. 2013). To identify local features that might discriminate CGI in the two different pausing classes, sequence motif elicitation was performed. The MEME motif finder (Machanick and Bailey 2011) was applied to 100 bp of sequence underlying the proximal and distal pause site for all three classes of genes. This identified G-rich sequences at the corresponding dominant Pol II pausing site for each class (**Figure 3A**); that is, proximally paused genes show G-enrichment near the TSS, while distally paused genes show a G-enrichment at the 3’ CGI boundary (**Figure 3A**). Genes in the silent class resemble that of the distally paused genes with G-rich sequences at the distal pause site, suggesting that, when active, these genes might exhibit a distally paused pattern similar to Class I genes.

The finding of G-rich coding strand sequences that correspond with the location of the dominant pause evoked an examination of GC ‘skew’. GC skew is a common feature of CGI promoters and is characterized by an excess of G versus C content on the coding strand (Ginno et al. 2012). Transcription through such regions results in the formation of R-loops generated by the stable pairing between the G-rich nascent RNA back to the C-rich template behind the progressing polymerase (Aguilera and Garcia-Muse 2012). This leaves the G-rich non-template DNA strand unpaired, which also has the potential to form G-quadruplexes (Lam et al. 2013; Shrestha et al. 2014). These secondary structures have been shown to promote DNA damage and/or translocations while also impeding transcription at certain genes (Aguilera and Garcia-Muse 2012). To investigate the possibility that GC skew influences Pol II pausing, GC skew was calculated for the region from the TSS to the downstream CGI edge and sorted by CGI size and predominant pausing site, again using the same sort order and class distinction as in **Figure 1C**. This demonstrated that GC skew does indeed correlate very closely with the dominant Pol II pausing site (**Figure 3B**). The compiled analysis of GC skew at all CGI-associated genes oriented to the TSS and scaled to the CGI showed that whereas there is generally positive skew downstream of the TSS as previously described (Ginno et al. 2013), the proximally paused genes have a sharp peak in GC skew located just downstream of the TSS and the distally paused genes have a sharp peak of GC skew at the CGI edge (**Figure 3C**). Moreover, sorting of all active CGI associated genes by the degree of GC skew at either the proximal or the distal site correlated with the degree of Pol II pausing at the same site, as indicated by the GRO-seq signal (**Figure 3D,E)**, and the interpolated pausing index (**Supplemental Figure 5A)**, but was independent of gene expression levels (**Supplemental Figure 5B,C**). Taken together these data indicate that Pol II pausing correlates more strongly with local GC skew than any of the chromatin modifications or trans-acting factors investigated.

Previous work by the Chedin group has shown a strong correlation between global GC skew, R-loop formation, and the lack of DNA methylation across CGI domains (Ginno et al. 2012; Ginno et al. 2013). Using a sequence-based algorithm, SkewR, that takes into account the degree, length, and direction of GC skew, as well as sequence composition (C+G content, CpG density) the group has classified CGI into 3 classes based on a predicted propensity for R-loop formation and direction of skew: 'Strong', 'Weak', and 'Reverse'. To determine the relationship between pausing classes and SkewR predicted propensity for R-loop formation, CGI in the distal, proximal and silent pausing class genes were annotated to GC skew classes. As expected, most of the CGI considered in this study were associated with "Strong" skew features (**Supplemental Figure 6A,B)**. Relative to the proximal class, the distally-paused class was somewhat enriched in CGI with "Strong" GC skew features, including slightly longer regions of positive GC-skew, as determined by SkewR peak length (median= 765bp vs. 822bp, proximal vs. distal, p=2.94e-8).), and slightly longer first exons, a feature previously correlated with the "Strong" skew class (median= 208bp vs. 237bp, proximal vs. distal, p=1.73e-8) (Ginno et al. 2013). CGI in the silent pausing class were depleted of "Strong" CGI and enriched in "Weak" and "Reverse" CGI. Consistent with this, the 'Reverse' skew class was also noted to be enriched in features of Polycomb-mediated repression (Ginno et al. 2013). Thus, although there was some overlap between pausing and skew classes, global GC skew features alone did not appear to be the primary determinant of pausing class. Rather the primary site of GC-skew, immediately downstream of the TSS or at the 3'CGI boundary, governs the predominant site of Pol II pausing within the promoter.

Next, we determined the relationship between R-loop formation and pausing class among CGI-associated promoters by examining DRIP (DNA:RNA IP)-seq data, an antibodybased approach that selectively captures RNA-DNA hybrids that has been adapted to massively parallel sequencing (Ginno et al. 2012). There was little DRIP-seq enrichment among CGI in the silent class, consistent with a relative lack of transcriptional activity (**Supplemental Figure 6D**). In contrast, both proximally-and distally-paused genes showed enrichment of R-loops over the CGI domain. Significantly, there was a greater enrichment of R-loops detected downstream of the TSS among the distally paused gene class that peaked at or near the 3’ edge of the CGI domain (**Supplemental Figure 6D**). These data indicate that there is a correlation between the stability and/or propensity to form R-loops and the propensity to undergo pausing at the distal pause site.

Several recent studies suggest that enhancers may function to regulate transcription in part by modulating Pol II pausing (Krumm et al. 1995; Brown et al. 1996; Core and Lis 2009; Zippo et al. 2009; Anand et al. 2013; Liu et al. 2013; Loven et al. 2013). We used chromatin conformation capture (3C) to probe the relationship between enhancer-promoter looping interactions and pausing class in MCF7 cells transiently exposed to estradiol. Upon estrogen stimulation, well-characterized distal enhancers for the *MYC*, *P2RY2*, and *SIAH2* genes are bound by the estrogen receptor alpha, resulting in the looping between the enhancer and promoter and rapid induction of gene expression (Fullwood et al. 2009; Wang et al. 2011; Li et al. 2013). The *MYC*, *P2RY2*, and *SIAH2* genes were chosen for this assay because they are rapidly induced in response to estrogen exposure, ensuring a direct transcriptional effect, and because their promoter-associated CGI are sufficiently large that the TSS and 3’ CGI edge can be readily resolved, allowing us to determine the spatial relationship between enhancer contacts and the paused Pol II (Hah et al. 2011; Danko et al. 2013; Hah et al. 2013). A high-resolution chromatin conformation capture (3C) assay was used to finely map estrogen-induced enhancer interactions at these loci in MCF7 cells. At the two distally paused genes, *MYC* and *SIAH2*, a 10 minute estrogen exposure induced a 5-10 fold increase in the efficiency of contact between the upstream enhancer and their respective promoters. Interestingly, the interaction efficiency was 2-3-fold greater at the distal edge of the CGI than surrounding regions, including the TSS (**Figure 4 A,B**). In contrast, at the proximally paused gene *P2RY2*, estrogen stimulated the interaction between the upstream enhancer and the TSS proximal region (**Figure 4C**). To assess the functional consequences of enhancer looping in this setting, we interrogated Gro-seq data obtained from estradiol stimulated MCF-7 cells (Hah, N. Cell 2011) to determine the influence of E2 stimulation on pausing indexes over time at the *MYC*, *SIAH2,* and *P2RY2* genes and calculated as described above (**Figure 3**). This analysis showed a ~2 fold decrease (2.2, 1.5, 2.0-fold, respectively) in pausing index at the corresponding distal (*MYC*, *SIAH2*) or proximal (*P2RY2*) pause point within ~40 minutes, and levelling off thereafter (data not shown).

The above data suggests that enhancer interactions may play a role in the regulation and/or stability of the Pol II pause not only at the proximal site but at the distal pause site as well. To examine the broader relationship between pausing class and enhancer-promoter interactions, we made use of the highest resolution (1 kb) Hi-C data available to date (Rao 2014 Cell) to examine the frequency of contacts between the proximal and distal pause sites of each CGI associated promoter, and its nearest enhancer. We defined enhancers as regions of overlap between H3K4me1/ H3K27Ac based on ChIP seq data from the same cell type from which the Hi-C data were derived (GM12878) (Consortium 2012) (Encode Project Consortium 2012). Given the resolution of the data, we limited the analysis to those CGI for which the TSS and the 3’ CGI edge are more than 1 kb apart (Proximal N=663 Distal N=475, Silent N=558), and among these, those that had one or more annotated contacts between the enhancer and each pause point. This resulted in a total of 1225 CGI promoter-enhancer pairs considered (Proximal=527, Distal=390, Silent=308). There was no significant difference in the distance from the promoter to the nearest enhancer between the proximal and distal pausing classes (**Supplemental Figure S7**). CGI promoters in the silent class were further from their nearest enhancer, which was not surprising considering the focus on 'active' H3K4me1/H3K27Ac marked enhancers, and the tendency for the silent class to be overrepresented in polycomb-marked genes, the bulk of which are likely silent in most differentiated tissues, including the GM12878 cells. We found that while the nearest enhancers associated CGI promoters in the proximal and silent classes were equally likely to contact the TSS as the 3'CGI edge, enhancers associated with CGI in the distal class showed a 1.3-fold greater propensity for contact at the distal pause site (p=.027, distal vs. proximal class genes, Mann-Whitney U-test) (**Supplemental Figure 7B**).

Several studies have implicated cohesin in long-range chromosomal contacts and enhancer-promoter looping (Kagey et al. 2010b).We therefore examined enrichment of the cohesin subunit Rad21 (MCF7 cells, (Consortium 2012)) in and around the TSS and 3' CGI edge of promoters in the three pausing classes. We found that while cohesin is enriched throughout the CGI in both active classes (proximal and distal) relative to the silent class, there was a skew towards more distal enrichment (decreased at the TSS increased at the 3' CGI edge) in the distally paused genes relative to the proximally paused genes (**Supplemental Figure 7C**).

Taken together, these data indicate that stable contacts are made between distal enhancers and multiple points along the CGI domain among active CGI-associated genes, with a tendency for distally paused genes to exhibit a skewed distribution towards the 3’ end of the CGI domain relative to proximally paused genes. Upon transcriptional activation, contacts at the dominant pause site appear to be selectively induced and /or stabilized, at least at the E2-regulated genes examined, suggesting a new role for enhancer-promoter contacts in Pol II pausing not only at the proximal pause, but the distal pause point as well.

Discussion:

This study provides evidence that at CGI-associated promoters Pol II encounters not one, but two major barriers to transcriptional elongation, one defined by the promoter proximal pause and a second that corresponds to the downstream boundary of the CGI domain. The relative degree to which each becomes rate-limiting is intrinsically determined by local sequence context. Both pausing events correlate with regions of high GC skew, a feature of sequences prone to the formation of secondary structures such as R-loops and G-quadruplexes (Aguilera and Garcia-Muse 2012; Ginno et al. 2012; Shrestha et al. 2014). Positive GC skew downstream of the TSS is a sequence feature of most CGI, and R-loops extending through this region have been detected at many CGI promoters where they are proposed to prevent DNA methylation (Ginno et al. 2012; Ginno et al. 2013). We find that even within the overall positive GC skew typical of most CGI, there are G-rich clusters that exhibit an even greater (2-3-fold) G-bias than the surrounding DNA, and that these correlate well, both in location and in magnitude, with the intensity of Pol II pausing. The hyperstability of the RNA: DNA duplexes formed at these sites in particular may tether the nascent transcript, impeding elongation beyond a certain distance or once a threshold level of negative supercoiling behind the progressing polymerase has been reached. Indeed, R-loop structures impede polymerase progression in a length and supercoiling dependent manner *in vitro* (Belotserkovskii et al. 2010) and promoter-proximal pausing has been correlated with local thermodynamic stability of the RNA:DNA duplex in *Drosophila* (Nechaev et al. 2010). Consistent with a tethering model, Pol II paused in close proximity to the G-rich sequences, not only at the promoter-proximal pause, but also at the distal pause, regardless of the distance from the TSS to the downstream CGI edge, and hence the overall length of the nascent transcript. Thus, even in the context of a nascent transcript well over 1 kb, G-rich stretches still have the capacity to hinder Pol II progression. A similar mechanism may be operative at the 3' ends of genes where Pol II pausing and R-loop formation over G-rich sequences downstream of the poly A signal are necessary for efficient transcription termination (Skourti-Stathaki et al. 2011; Skourti-Stathaki et al. 2014).

Similar G-rich clusters interspersed throughout the immunoglobulin class switch locus are proposed to serve as points of R-loop initiation, with the R-loop spreading laterally from the point of the most stable (highest G-bias) contact (Zhang et al. 2014). Consistent with this idea, the distally paused genes were enriched in RNA:DNA hybrids that peaked near the 3'CGI edge relative to proximally-paused genes. It is possible that the additional stability afforded by the intertwining of a longer nascent transcript "tail" with the template DNA (eg. long R-loop) at distally paused genes might obviate the need for additional protein factors to enforce the paused state. Indeed, whereas the protein factors known to enforce (NELF, DSIF) or to relieve (BRD4 and P-TEFb) promoter-proximal pausing are enriched at the proximal site across active CGI of both classes (**Supplemental Figure 3B**), there was little enrichment of these factors at the distal pausing site in either class, suggesting that the mechanisms involved in proximal and distal pausing and release may differ. A paused state reinforced by additional trans-acting factors versus one driven by sequence and physical constraints alone might explain the difference in apparent intensities between the proximal versus the distal pause (**Supplemental Figure 8).** Interestingly, classes of genes exhibiting a more proximal versus distal pause have also recently been reported in *Drosophila,* which lack CpG islands (Kwak et al. 2013). The strong correlation between the position of pausing and GC skew is also readily evident in the *Drosophila* data (**Supplemental Figure 9**), suggesting that the relationship between this sequence-based feature and Pol II pausing is evolutionarily conserved, and may predate the appearance of CGI.

We demonstrate that enhancers contact the predominant pausing site in CGI, suggesting a role for enhancers in regulating Pol II pausing not only at the proximal pause, but also at the CGI edge. The role of active enhancers in mediating promoter-proximal pause release is incompletely understood, but has been attributed to the delivery of the P-TEFb complex to the promoter, enhancer-mediated liberation of P-TEFb from local HEXIM1-mediated sequestration, and most recently, a mechanism involving competition for promoter-bound NELF by enhancer-derived eRNAs (Anand et al. 2013; Liu et al. 2013; Loven et al. 2013; Schaukowitch et al. 2014). It is tempting to speculate that at distally paused genes, the enhancer might bring in factors capable of resolving R-loops or other physical constraints (eg. supercoiling). To this end, inhibition or downregulation of Type I topoisomerases leads to an accumulation of R-loops at the *MYC* locus (Yang et al. 2014) or ribosomal RNA genes (El Hage et al. 2010; Marinello et al. 2013). How enhancers might be selectively directed to the promoter versus the CGI edge is unclear, but recent evidence showing that substantial fraction of long-range looping interactions are stable across cell types and developmental windows, and precede signal induced gene activation (Jin et al. 2013; Li et al. 2013; Ghavi-Helm et al. 2014) suggests that at least one component of enhancer activity (looping) may be linked to underlying sequence features.

Our findings highlight the importance of considering the CGI and its embedded TSS as a discrete chromatin domain whose structure has implications not only for transcriptional initiation but also elongation. While it is well accepted that the maintenance of an open chromatin conformation at CGI facilitates promoter access and Pol II loading (Deaton and Bird 2011), our data suggest that the CGI boundaries act, to varying degrees, as a natural barrier to transcriptional elongation, in both the sense (productive) and antisense (divergent) directions (see **Figure 1**, **Supplemental Figure 2B**). Previous work by the Chedin group has shown GC skew is a common feature of CpG islands and that transcription through these regions, rather than transcription per se, is important for protecting CpG islands from DNA methylation in transfected plasmids, implying a role for R-loop formation in this process (Ginno et al. 2012; Ginno et al. 2013). We have previously shown that Pol II, even in the paused state, can protect CpG islands from *de novo* methylation after drug-induced demethylation (Kagey et al. 2010a). The evolutionary preservation of CpG density in CGI has been attributed to the absence of DNA methylation in the germline, and hence a reduced rate of spontaneous meC to T transition mutations in these regions relative to the rest of the genome. The co-evolution of G-C strand asymmetry in the CGI domain implies a transcriptionally-dependent event that drives the preferential loss of non-CpG C's on the coding strand (or Gs on the template strand). The preservation of highest GC-skew in CGI relative to the genome-wide average (Ginno et al. 2013) and in particular at the positions where RNA Pol II is paused (this study) suggests a role for Pol II residency time in promoting this event. Together these data add to a growing body of evidence supporting the idea that divergent transcription, GC skew/ R-loop formation, and Pol II pausing are inextricably linked, and conspire to maintain the unique epigenetic environment of CGI domains.

Methods:

Datasets used in this study:

The following datasets were used in this study: MCF7 cells; ChIP-seq of Pol II (GSM365929) (Welboren et al. 2009), phospho-S5 Pol II (GSM588577)(Joseph et al. 2010), H3K4me3 (GSM945269) (2012), H3K4me2 (GSM822391) (He et al. 2012), H3K9ac (GSM588573) (Joseph et al. 2010), H3K9me1 (GSM945857) (2012), H3K14ac (GSM588575) (Joseph et al. 2010), H3K27ac (GSM946850) (Frietze et al. 2012), MYC (GSM1006877) (2012), and Rad21 (GSM101079) (2012)**;** MeDIP-seq (DRX000030) (Ruike et al. 2010); and GRO-seq (GSM1014637)(Hah et al. 2011). IMR90 cells; GRO-seq (GSM340901) (Core et al. 2008). MCF10A cells; GRO-seq (ERX016683) (Kim et al. 2013). CD4+ T cells; ChIP-seq of Pol IIS5 (GSM1022949) (Zhang et al. 2012) and BRD4 (GSM823378) (Zhang et al. 2012). HeLa cells; ChIP-seq of NELFA (GSM1280296)(Liu et al. 2014) and the SUPT5H component of DSIF (GSM1280295)(Liu et al. 2014). K562 cells; ChIP-seq of CCNT2 (GSM935547) (2012). GM12878 cells; H3K4me1 and H3K27Ac ChIP-seq peak files (GSM733771, GSM733772) (2012), and Hi-C contact matrices at 1kb resolution (GSE63525) (Rao et al. 2014) Ntera cells; DRIP-seq (SRX113813) (Ginno et al. 2012); Drosophila; Pro-seq (GSE42117) (Kwak et al. 2013).

Heat map representations of ChIP-seq, MeDIP-seq, and GRO-seq tag densities were generated by summing the total number of tags in 20 base pair bins +/- 3 kb to either side of the TSS or the midpoint of the CGI domain and visualized with Java Treeview v.1.1.6

Heat maps, CGI Scaling and Genome-wide averages:

CpG islands were defined by the UCSC criteria (Hg19; April 27, 2009) (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz). Heat map representations of ChIP-seq, MeDIP-seq, and GRO-seq tag densities were generated by summing the total number of tags in 20 base pair bins +/- 3 kb to either side of the TSS or the midpoint of the CGI domain and visualized with Java Treeview v.1.1.6. To compare CGI features, CGI associated TSS were oriented to the direction of transcription and the distance from the TSS to the 5’ and 3’ edge of the CGI were independently scaled. It should be noted that as a run-on technique, GRO-seq tags will map approximately 100 bases downstream of the Pol II position. As a consequence, the summit of GRO-seq enrichment at the proximal pause maps to ~+140 to +240 relative to the TSS, and is offset from the peak of Pol II ChIP enrichment (~ +40 to +140 from the TSS) by about 100bp. Likewise, the GRO-seq enrichment at the CGI boundary is ~+40 to +140 of the 3’ edge of the CGI on average. To account for this offset, 200 bp was added to the length of the CGI for GRO-seq scaling purposes in order to fully account for the second peak. The length of each half of the CpG island was divided into 50 bins and the average number of tags from each bin was compiled for all promoter-associated CGI or the indicated class using an in-house perl script. Information for a constant (unscaled) 800bp was included on each side of the CGI for comparison.

Proximal and Distal pausing indices:

The pausing index was defined as the total number of GRO-seq tags for the 100 bases spanning the proximal pause (+140 to +240 relative to TSS) or the distal pause (+20 to+120 relative to 3’CGI edge) divided by the average number of reads per 100 bases across the gene body, which was defined as the region from +200 bases downstream of the 3’ CGI edge to the transcription end site (TES). TSS associated CGI were parsed into the dominant pausing class by comparing the ratio of the proximal to distal pausing index. Those CGI with a value less than 1 were placed in the distal class, while those with a value greater than 1 were placed in the proximal class. CGI with no tags at either pause were placed in the silent class.

GC Skew and Gene Expression:

GC skew (G-C/G+C) was calculated in 20 bp bins across the human Hg19 genome and the skew +/- 3kb upstream and downstream to generate heatmaps sorted by CGI size of the three classes. For sorting of GC skew, all CGI promoters were sorted by decreasing GC skew for the 100 bases of enrichment (the 100 bases immediately upstream from CGI edge for the distal sort, and +40 to +140 from TSS for the proximal sort). For scaling GC skew, a 20 bp sliding window moving in 1 bp increments was calculated and scaled into a fixed number of bins (n=50) independently from the TSS to the upstream and downstream CGI edges. The average of each bin across all genes was plotted for each class. Gene expression was determined from GRO-seq “sense” tag density as RPKM for the gene body region defined as +200 bp downstream of the 3’ edge of the CGI to the TES. GC skew (G-C/G+C) was calculated across the *Drosophila* genome (dm5.22) as described above and the average skew for each bin centered on the TSS +/- was plotted for all genes in each pausing class (proximal, distal) as defined in Kwak et al. (Kwak et al. 2013).

Chromatin Conformation Capture (3C):

Experiments were carried out according to Hagege et al. (Hagege et al. 2007) with slight modification. Briefly, 15cm plates containing 1 X 10^6 MCF-7 cells were grown in phenol-red free media supplemented with 10% charcoal stripped serum for 4 days to deplete estrogen. Cells were treated with 100 nM estradiol or vehicle (ethanol) for 10 min. at 37°C followed by crosslinking in 1% formaldehyde in the media for 10 min. Glycine was added to a final concentration of 0.125M for 10 min. at room temperature to stop the reaction. Cells were washed twice with PBS and scraped into cold lysis buffer (10mM Tris pH 8.0, 10mM NaCl, 0.2% NP-40), transferred to a conical tube and placed on ice for 15 min. with occasional mixing. Nuclei were pelleted at 400g for 5 min. and washed once in 1X restriction buffer. Restriction digests were carried out in 1X NEB restriction buffer 3.1 or CutSmart, 1% TX-100, and 400 U each restriction enzyme per sample and incubated overnight at 37°C with rocking. Enzymes were inactivated with 1.25% SDS at 65°C for 25 min. Samples were then ligated in 1X NEB ligation buffer, 1% TX-100, 200 ˪g BSA, 3000 U of ligase (NEB M0202) in a total of 7.5 mL for 4h at 16°C. DNA was isolated by proteinase K (500 mg) digestion overnight at 65°C, followed by phenol:chloroform extraction and ethanol precipitation. Ligation efficiency between the distal enhancer and various restriction fragments was interrogated by quantitative realtime PCR using Taqman probes. A constant forward primer and Taqman probe anchored at the distal enhancer fragment were coupled with forward primers designed against selected restriction fragments across the region and every restriction fragment within the CGI. The data is represented as the average fold change in estradiol induced samples relative to vehicle-only controls from three independent biological experiments assayed in triplicate. Primers used for 3C analysis are listed in Supplemental Table I.

Restriction enzymes used for the analysis of the *MYC* and *SIAH2* loci were PstI (NEB #R0140) and NsiI (NEB #R0127); *P2RY2* was XmaI (NEB #R0180) and BsaWI (NEB #R0567). The *P2RY2* digest was modified slightly by sequential digestion first with BsaWI at 60°C for 30 minutes followed by the addition of XmaI and a second 400 U of BsaWI and incubation at 37°C overnight. This led to complete digestion of test BsaWI fragments (data not shown).

Hi-C Enhancer Contact Analyses

ENCODE Chip-Seq data from GM12878 cells, GSM733771, GSM733772) was used to define active enhancers as regions of overlap between H3K4me1 and H3K27Ac peak files(Consortium 2012) .This identified 52,422 genomic regions with a median width of 827bp. Each CGI promoter was then annotated to the nearest putative enhancer region to its TSS. Given the resolution of the data, the analysis was limited to only those genes where the TSS and 3'CGI edge are at least 1 kb apart (Proximal N=663 Distal N=475, Silent N=558), and among those, those that had one or more contacts between each pause site and the nearest enhancer (Proximal N=527, Distal N=390, Silent N=308). Hi-C contact matrices from GM12878 cells at 1kb resolution (MAPQ score >30) were obtained from Rao et al.(Rao et al. 2014)(GSE63525), and were used to determine the frequency of contacts between each pausing site and the nearest enhancer. Contacts between the proximal (TSS+200bp) and distal (3'CGI edge +/- 100bp) and the nearest enhancer were counted, and the log2 of the ratio of distal pause site contacts to proximal pause site contacts was calculated.

R-loop, SkewR Class Analysis

SkewR peaks and TSS class assignments are available from https://www.mcb.ucdavis.edu/faculty-labs/chedin/Resources.html (Ginno et al. 2012; Ginno et al. 2013) For Skew R peak length analysis, TSS from genes in each pausing class were intersected with SkewR peaks (low stringency). Approximately 80% of the TSS in each class were found within a SkewR peak (Proximal N=4348, Distal N=2730, Silent N=2641). To annotate CGI to a skew ‘class’, CGI were matched to the nearest TSS within 1kb for which a SkewR class assignment was available (Ginno et al. 2013)(Proximal N=5601, Distal N=3490, Silent N=2461).

DRIP-Seq data tag densities were calculated from (SRX113813) (Ginno et al. 2013). For scaling purposes, CGI associated TSS were oriented to the direction of transcription and the distance from the TSS to the 5' and 3' edge of the CGI were independently scaled into 40 bins each. An additional fixed (unscaled) distance of 800 bp (40x20 bp bins) was included to each side for comparison. After removing duplicate reads the average tag densities across all genes in each pausing class was calculated using custom R scripts.

Motif Elicitation

MEME-ChIP (http://meme-suite.org/tools/meme-chip) was performed on the 100 bp sequence underlying the proximal pause and the distal pause for the promoters in each pausing class (Distal, Proximal, Silent) with the following options: Background sequence model, 1st order, scan given strand only. All other MEME options were set to the default values.

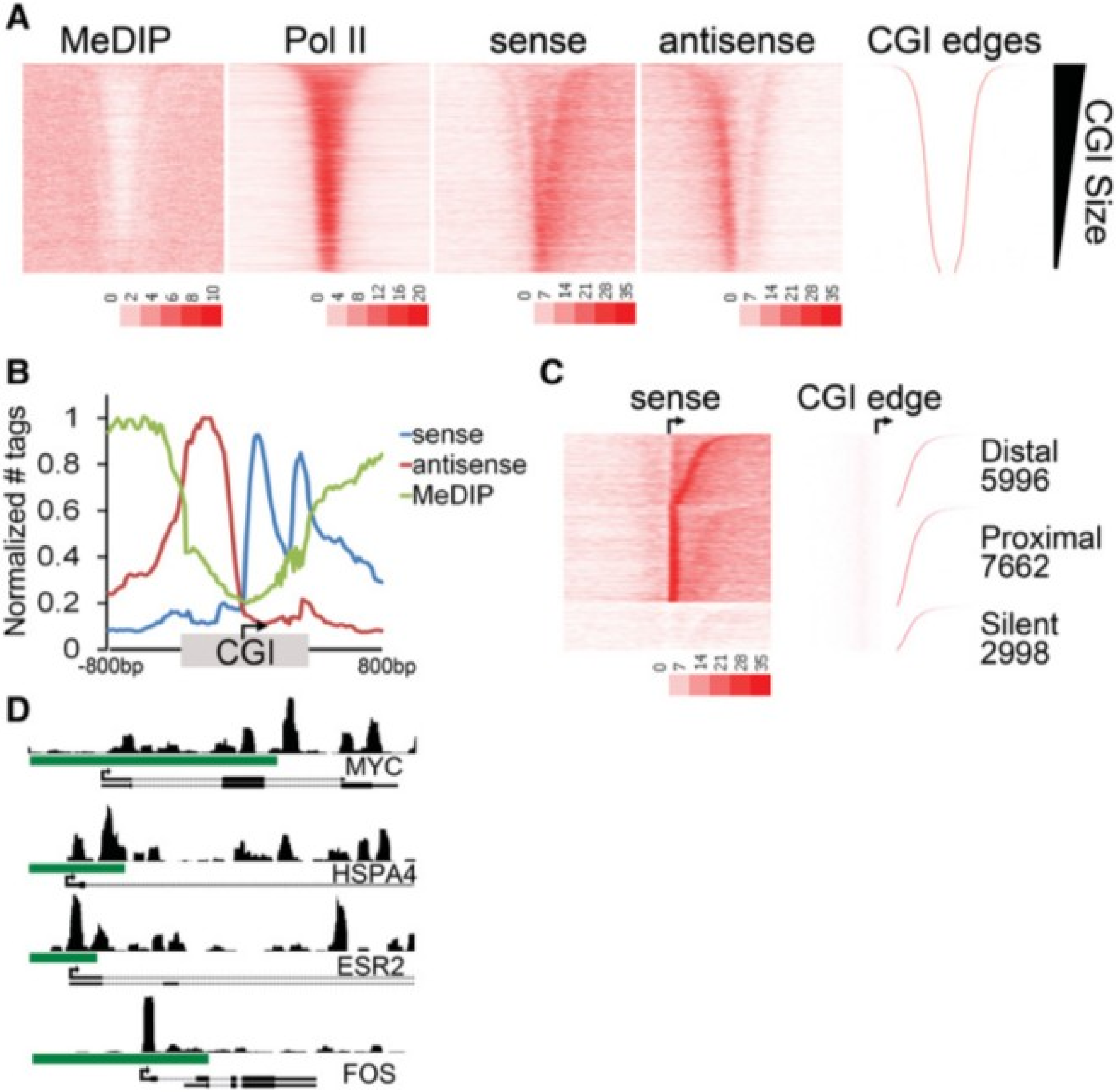
Acknowledgments: We thank the authors who have publically submitted the datasets used in this study, and Dr. Paul Wade for thoughtful critique of the manuscript and helpful advice. This work was supported by NIH grants 2R01 CA077337 and 5R01 CA132065 (to PMV), and 1F31 CA186676 (to JSKB).

Disclosure Declaration: The authors have nothing to disclose.

Figure Legends:

Figure 1. A second (distal) Pol II pause at the CGI shore.

(A) Heatmap representation of DNA methylation (MeDIP-seq), total Pol II ChIP-seq, and GRO-seq sense (plus strand) and antisense (minus strand) tag density is plotted for +/- 3kb around the midpoint of the CGI and sorted by CGI size (CGI associated promoters; n=16,657). The upstream and downstream boundaries of the CGI domain (right panel) is shown for comparison (B) Average tag densities of nascent transcripts (GRO-seq; sense, antisense) and DNA methylation (MeDIP-seq) across CGI associated promoters. Promoters were oriented to the direction of transcription and the distance from the TSS to the upstream and downstream CGI edge were independently scaled and anchored to the TSS (arrow). An additional 800 bp to either side of the CGI (unscaled) is included. Data are normalized between datasets by setting the maximum tags per 20 bp bin within each dataset to 1. (C) The relative GRO-seq tag density for the 100 bp under the proximal peak versus the 100 bp at the CGI edge was used to parse genes into proximal or distal pausing classes. Promoters with no tags in either region were considered silent. CGI in each class were sorted by the distance from the TSS to the downstream CGI edge (indicated to the right). (D) Browser image of MCF-7 GRO-seq sense tags covering an 8 kb window surrounding the promoter regions of *MYC*, *HSPA4*, *ESR2*, and *FOS* (Green bar = CGI).


A
MeDIP
Pol II
sense
antisense
CGI edges
CGI Size
0 2 4 6 8 10
0 4 8 12 16 20
0 7 14 21 28 35
0 7 14 21 28 35
B
Normalized # tags
1
0.8
0.6
0.4
0.2
0
-800bp
CGI
800bp
sense
antisense
MeDIP
C
sense
CGI edge
Distal
5996
Proximal
7662
Silent
2998
0 7 14 21 28 35
D
MYC
HSPA4
ESR2
FOS

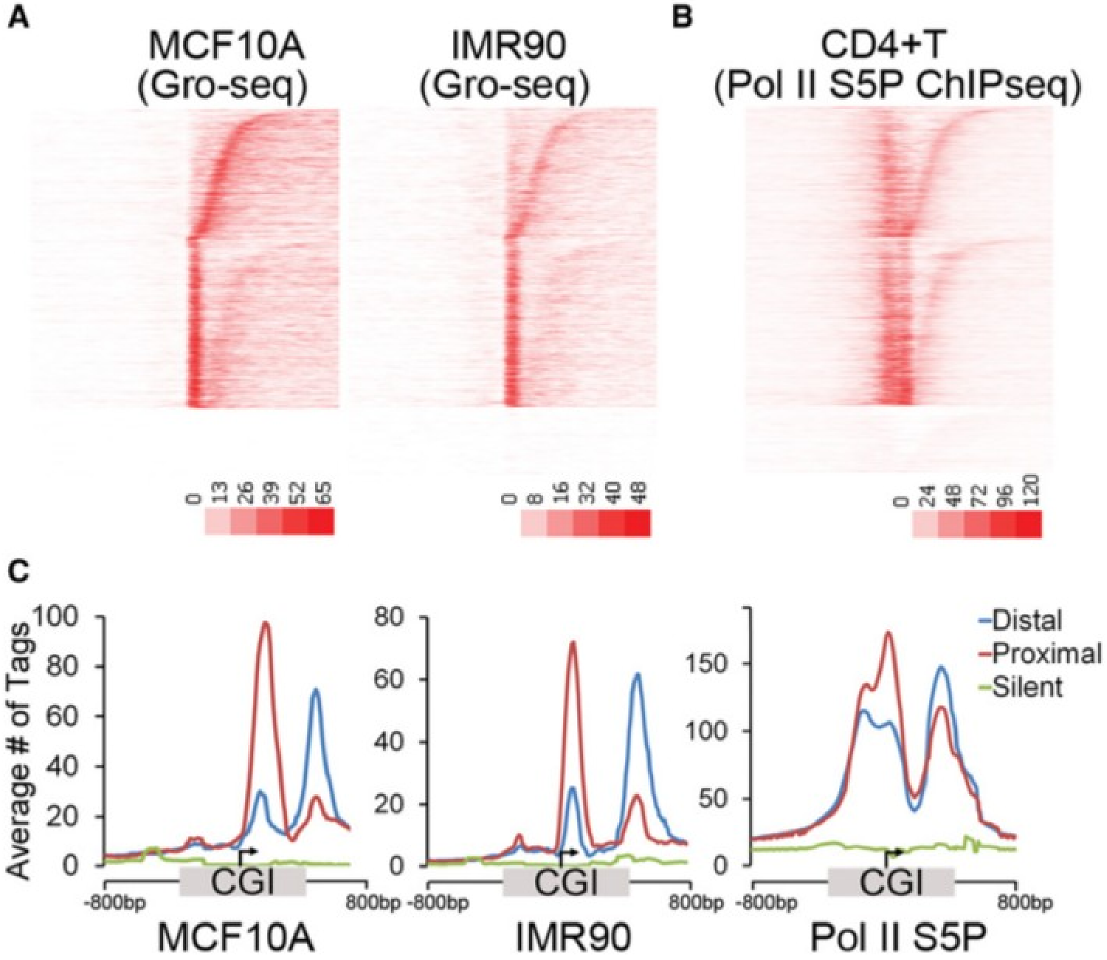


Figure 2. Predominant Pol II pausing class is conserved across cell types.

(A) Heat map representation of the GRO-seq sense tag density from MCF10A and IMR90 cells. CGI promoters are oriented to transcription and sorted within each class by the distance from the TSS to the downstream CGI edge using the same sort order as Figure 1C. (B) Pol II (S5 phosphorylated) ChIP-seq tag density from CD4+ T cells oriented and sorted as in Figure 1C. (C) Average tag densities of nascent transcripts (GRO-seq sense) and phospho-S5-Pol II for promoters in the three pausing classes shown in A and B. CGI associated promoters were oriented to transcription and the distance from the TSS to the upstream and downstream CGI edge independently scaled and anchored to the TSS (arrow). The average tags per 20 bp bin for 800 bp to either side of the CGI (unscaled) is included.

Figure 3. Pol II pausing correlates with GC skew.

MEME-ChIP was performed on the 100 bp sequence underlying the proximal pause and the distal pause for the promoters in each pausing class (Distal, Proximal, Silent). An enrichment of G-rich sequences correlates with the predominant Pol II pause point for each class. (B) Heat map representation of GC skew. The degree of GC skew was calculated in 20 base pair bins. CGI promoters from the three classes are oriented and sorted by the distance from the TSS to the downstream CGI edge using the same sort order as Figure 1C. (C) Average GC skew across the three different pausing classes. CGI associated promoters were oriented and the distance from the TSS to the upstream and downstream CGI edge were independently scaled and anchored to the TSS (arrow). An additional 800 bases to either side of the CGI (unscaled) is included. (D, E) MCF-7 cell GRO-seq sense tag density around the TSS (arrow) or the downstream edge of the CGI (+/- 3kb). All CGI promoters were sorted by decreasing GC skew for the 100 bases underlying the distal pause (D) or the proximal pause (E).

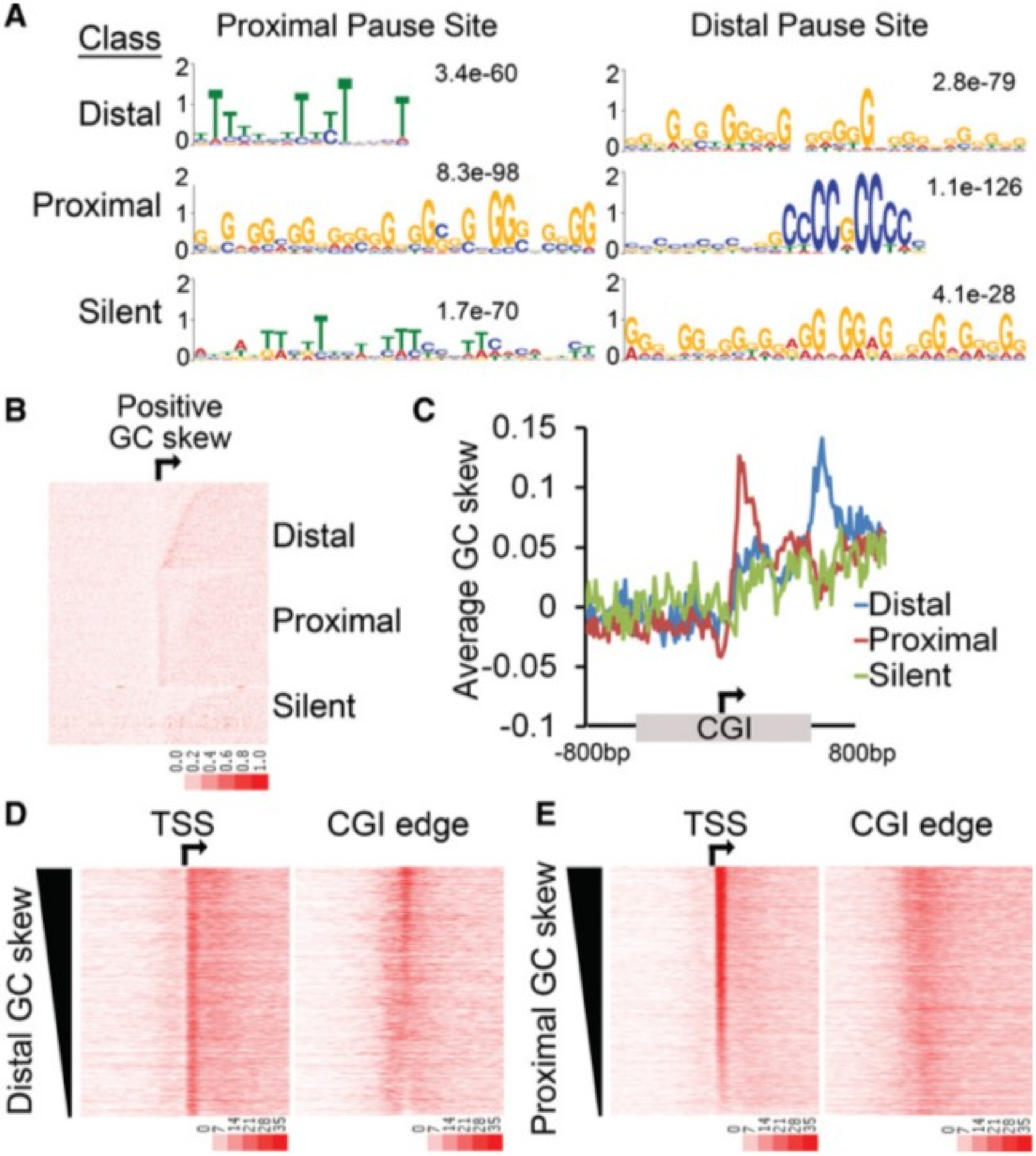
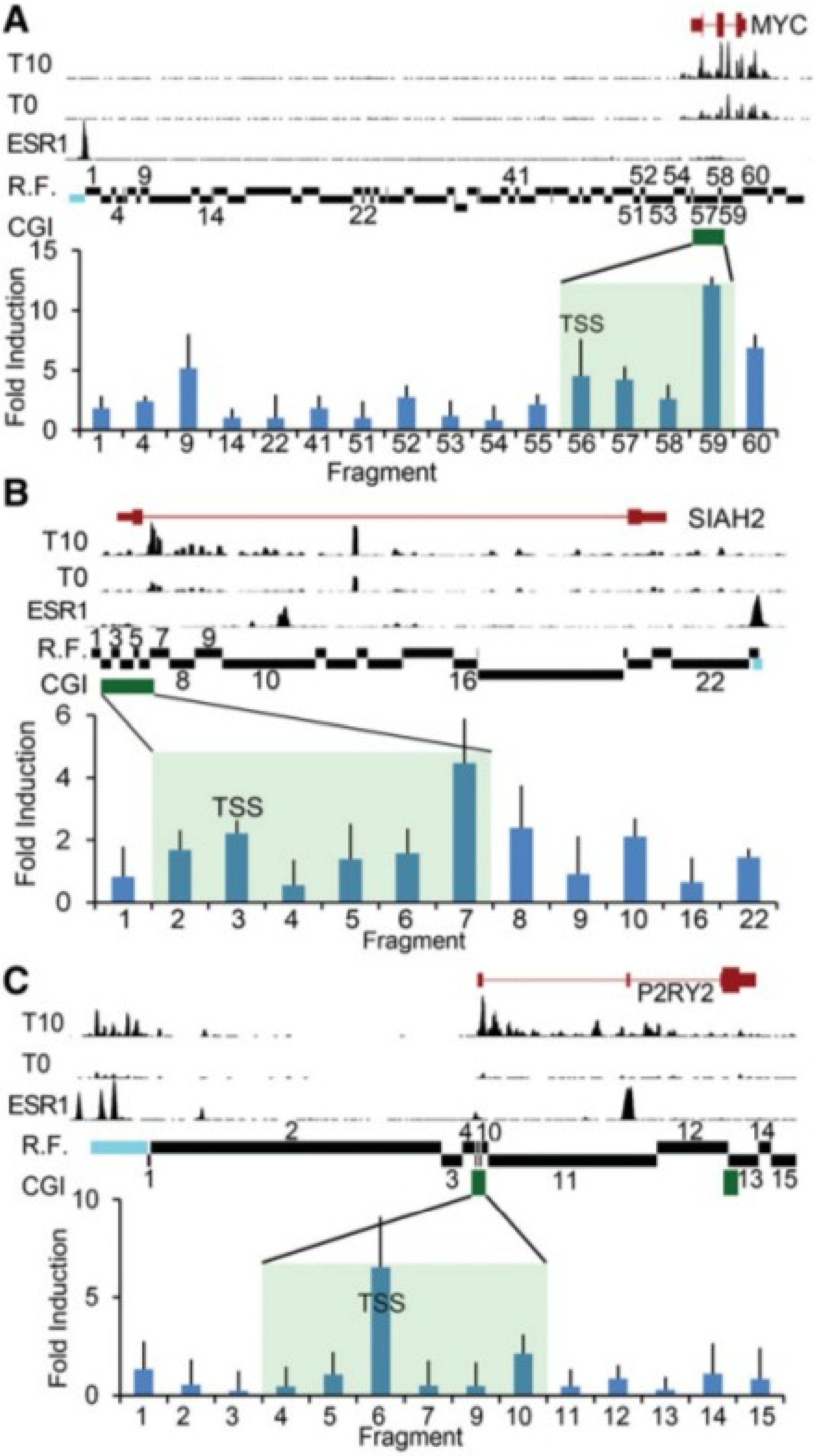

A
Class
Proximal Pause Site
Distal Pause Site
Distal
3.4e-60
2.8e-79
Proximal
8.3e-98
1.1e-126
Silent
1.7e-70
4.1e-28
B
Positive GC skew
Distal
Proximal
Silent
0.0 0.2 0.4 0.6 0.8 1.0
C
Average GC skew
0.15
0.1
0.05
0
-0.05
-0.1
-800bp
800bp
CGI
Distal
Proximal
Silent
D
TSS
CGI edge
Distal GC skew
0 7 14 21 28 35
E
TSS
CGI edge
Proximal GC skew
0 7 14 21 28 35

Figure 4. Distal enhancer interactions correlate with pausing class.

A chromatin conformation capture (3C) assay was performed to investigate the interaction between known estrogen-bound enhancer elements upstream of the (A) *MYC* (shown is chr8:128,679,000-128,763,000), (B) *SIAH2* (chr3:150,455,000-150,483,000), and (C) *P2RY2* (chr11:72,903,000-72,950,000) loci. The relative position of the CGI (green bar, shaded region) and the fragment containing the TSS are indicated. Estrogen-depleted MCF-7 cells were induced with 100 nM estradiol or vehicle (ethanol) for 10 minutes followed by cross linking, restriction digestion and ligation. An anchor probe was designed against the known estrogen receptor bound enhancer (blue) and tested for ligation with the indicated restriction fragments (R.F.) by qPCR. Data are reported as mean +/- s.d. of the fold-induction of E2 induced interaction relative to uninduced from three independent experiments assayed in triplicate. Shown for comparison are GRO-seq sense strand data from MCF-7 cells depleted of estrogen for 3 days (T0) and induced with 100 nM estradiol for 10 minutes (T10)(Hah et al. 2011). For GRO-seq tracks, Y axis scale (total tag count every 10 bases): MYC = 1300, SIAH2 = 250, P2RY2 = 225). Shown for comparison is estrogen receptor alpha (ESR1) ChIP-seq data derived from GSM594602. For the ESR1 ChIP-seq track, Y axis scale (total tag count every 10 bp): *MYC* = 150, *SIAH2* = 154, *P2RY2* = 324). Data demonstrate estrogen-induced ESR1 binding at the enhancer and transcriptional activity at both the promoter and enhancer.


A
MYC
T10
T0
ESR1
R.F.
1 9 41 52 54 58 60
4 14 22 51 53 5759
CGI
15
10
5
0
Fold Induction
TSS
1 4 9 14 22 41 51 52 53 54 55 56 57 58 59 60
Fragment
B
SIAH2
T10
T0
ESR1
R.F.
1 3 5 7 9
8 10 16 22
CGI
6
4
2
0
Fold Induction
TSS
1 2 3 4 5 6 7 8 9 10 16 22
Fragment
C
P2RY2
T10
T0
ESR1
R.F.
2 4 10 12 14
1 3 11 13 15
CGI
10
5
0
Fold Induction
TSS
1 2 3 4 5 6 7 9 10 11 12 13 14 15
Fragment

References:

The Encode Project Consortium 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**: 720-731.

Aguilera A, Garcia-Muse T. 2012. R loops: from transcription byproducts to threats to genome stability. *Mol Cell* **46**: 115-124.

Anand P, Brown JD, Lin CY, Qi J, Zhang R, Artero PC, Alaiti MA, Bullard J, Alazem K, Margulies KB et al. 2013. BET bromodomains mediate transcriptional pause release in heart failure. *Cell* **154**: 569-582.

Belotserkovskii BP, Liu R, Tornaletti S, Krasilnikova MM, Mirkin SM, Hanawalt PC. 2010. Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. *Proc Natl Acad Sci U S A* **107**: 12816-12821.

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* **2**: e26.

Brown SA, Imbalzano AN, Kingston RE. 1996. Activator-dependent regulation of transcriptional pausing on nucleosomal templates. *Genes Dev* **10**: 1479-1490.

Core LJ, Lis JT. 2009. Paused Pol II captures enhancer activity and acts as a potent insulator. *Genes Dev* **23**: 1606-1612.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**: 1845-1848.

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**: 212-222.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010-1022.

Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi J, Jacobs HM, Kastritis E, Gilpatrick T, Paranal RM, Qi J et al. 2011. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* **146**: 904-917.

El Hage A, French SL, Beyer AL, Tollervey D. 2010. Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev* **24**: 1546-1558.

Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. 2003. Predicting aberrant CpG island methylation. *Proc Natl Acad Sci U S A* **100**: 12253-12258.

Frietze S, Wang R, Yao L, Tak YG, Ye Z, Gaddis M, Witt H, Farnham PJ, Jin VX. 2012. Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol* **13**: R52.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**: 58-64.

Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EE. 2014. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**: 96-100.

Gilchrist DA, Dos Santos G, Fargo DC, Xie B, Gao Y, Li L, Adelman K. 2010. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**: 540-551.

Ginno PA, Lim YW, Lott PL, Korf I, Chedin F. 2013. GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* **23**: 1590-1600.

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell* **45**: 814-825.

Hagege H, Klous P, Braem C, Splinter E, Dekker J, Cathala G, de Laat W, Forne T. 2007. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* **2**: 1722-1733.

Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL. 2011. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**: 622-634.

Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. 2013. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**: 1210-1223.

Hashimoto H, Vertino PM, Cheng X. 2010. Molecular coupling of DNA methylation and histone methylation. *Epigenomics* **2**: 657-669.

He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. 2012. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res* **22**: 1015-1025.

Jang MK, Mochizuki K, Zhou M, Jeong H-S, Brady JN, Ozato K. 2005. The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* **19**: 523-534.

Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X. 2007. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**(7159): 248-251.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B.

2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290-294.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**: 484-492.

Joseph R, Orlov YL, Huss M, Sun W, Kong SL, Ukil L, Pan YF, Li G, Lim M, Thomsen JS et al. 2010. Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Mol Syst Biol* **6**: 456.

Kagey JD, Kapoor-Vazirani P, McCabe MT, Powell DR, Vertino PM. 2010a. Long-term stability of demethylation after transient exposure to 5-aza-2′-deoxycytidine correlates with sustained RNA polymerase II occupancy. *Mol Cancer Res* **8**: 1048-1059.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS. 2010b. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**(7314): 430-435.

Kanno T, Kanno Y, LeRoy G, Campos E, Sun H-W, Brooks SR, Vahedi G, Heightman TD, Garcia BA, Reinberg D. 2014. BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. *Nature Stru Mol Biol.***21**: 1047-1057.

Kim YJ, Greer CB, Cecchini KR, Harris LN, Tuck DP, Kim TH. 2013. HDAC inhibitors induce transcriptional repression of high copy number genes in breast cancer through elongation blockade. *Oncogene* **32**: 2828-2835.

Krumm A, Hickey LB, Groudine M. 1995. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev* **9**: 559-572.

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**: 950-953.

Kwak H, Lis JT. 2013. Control of transcriptional elongation. *Annual Rev.Genet.* **47**: 483.

Lam EY, Beraldi D, Tannahill D, Balasubramanian S. 2013. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat Commun* **4**: 1796.

Lee BK, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, Ni Y, Birney E, Lieb JD, Furey TS et al. 2012. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res* **22**: 9-24.

Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X et al. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**: 516-520.

Liu P, Xiang Y, Fujinaga K, Bartholomeeusen K, Nilson K, Price DH, Peterlin BM. 2014. Release of P-TEFb from 7SK snRNP Activates HEXIM1 Transcription. *J Biol Chem* 289: 9918-25...

Liu W, Ma Q, Wong K, Li W, Ohgi K, Zhang J, Aggarwal AK, Rosenfeld MG. 2013. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155**: 1581-1595.

Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. 2013. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**(2): 320-334.

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**: 1696-1697.

Marinello J, Chillemi G, Bueno S, Manzo SG, Capranico G. 2013. Antisense transcripts enhanced by camptothecin at divergent CpG-island promoters associated with bursts of topoisomerase I-DNA cleavage complex and R-loop formation. *Nucleic Acids Res* **41**: 10110-10123.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**: 335-338.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc regulates transcriptional pause release. *Cell* **141**: 432-445.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.

Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G. 2010. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics* **11**: 137.

Schaukowitch K, Joo J-Y, Liu X, Watts JK, Martinez C, Kim T-K. 2014. Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Mol Cell* **56**: 29-42.

Shrestha P, Xiao S, Dhakal S, Tan Z, Mao H. 2014. Nascent RNA transcripts facilitate the formation of G-quadruplexes. *Nucleic Acids Res* **42**: 7236-7246.

Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ. 2014. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* 516:436-9.

Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* **42**: 794-805.

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**: 1082-1086.

Wang C, Mayer JA, Mazumdar A, Fertuck K, Kim H, Brown M, Brown PH. 2011. Estrogen induces c-myc gene expression via an upstream enhancer activated by the estrogen receptor and the AP-1 transcription factor. *Mol Endocrinol* **25**: 1527-1538.

Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG. 2009. ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J* **28**: 1418-1428.

Yang Y, McBride KM, Hensley S, Lu Y, Chedin F, Bedford MT. 2014. Arginine methylation facilitates the recruitment of TOP3B to chromatin to prevent R loop accumulation. *Mol Cell* **53**: 484-497.

Yang Z, Yik JH, Chen R, He N, Jang MK, Ozato K, Zhou Q. 2005. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell* **19**: 535-545.

Zhang W, Prakash C, Sum C, Gong Y, Li Y, Kwok JJ, Thiessen N, Pettersson S, Jones SJ, Knapp S. 2012. Bromodomain-containing protein 4 (BRD4) regulates RNA polymerase II serine 2 phosphorylation in human CD4+ T cells. *J Biol Chem* **287**: 43137-43155.

Zhang ZZ, Pannunzio NR, Han L, Hsieh CL, Yu K, Lieber MR. 2014. The Strength of an Ig Switch Region Is Determined by Its Ability to Drive R Loop Formation and Its Number of WGCW Sites. *Cell Rep* **8**: 557-569.

Zippo A, Serafini R, Rocchigiani M, Pennacchini S, Krepelova A, Oliviero S. 2009. Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell* 138: 1122-1136.

Supplemental Data and Methods

Supplemental Methods:

GSEA/ GO Analysis

Gene sets from each pausing class were analyzed for overlap with curated data sets (C2, C4, C6, C7, H, MF) in MSigDB using the web interface available at http://www.broadinstitute.org/gsea/msigdb/ (Subramanian a. et al. 2005) and for functional annotation using the DAVID Bioinformatics Resource (http://david.abcc.ncifcrf.gov) (Dennis et al. 2003). CGI-associated TSSs in each class were annotated to the corresponding gene. Input gene lists (symbols) were created from genes ranked by pausing index (eg. Top 20% by PI) within each class or by the ratio of Proximal to Distal (or Distal to Proximal) pausing indexes (eg. Top 1000 ratio Prox:Dist, etc). For the silent class, all genes for which the average reads/kb in the gene body >0 was used as input.

Supplemental References:

Dennis G, Sherman, BT, Hosack, DA, Yang J, Baseler BW, Lane HC, Lempicki, RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4: P3.

Ginno PA, Lim YW, Lott PL, Korf I, Chedin F. 2013. GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res 23: 1590-1600.

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45: 814-825.

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339: 950-953.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A. 102:15545-50.

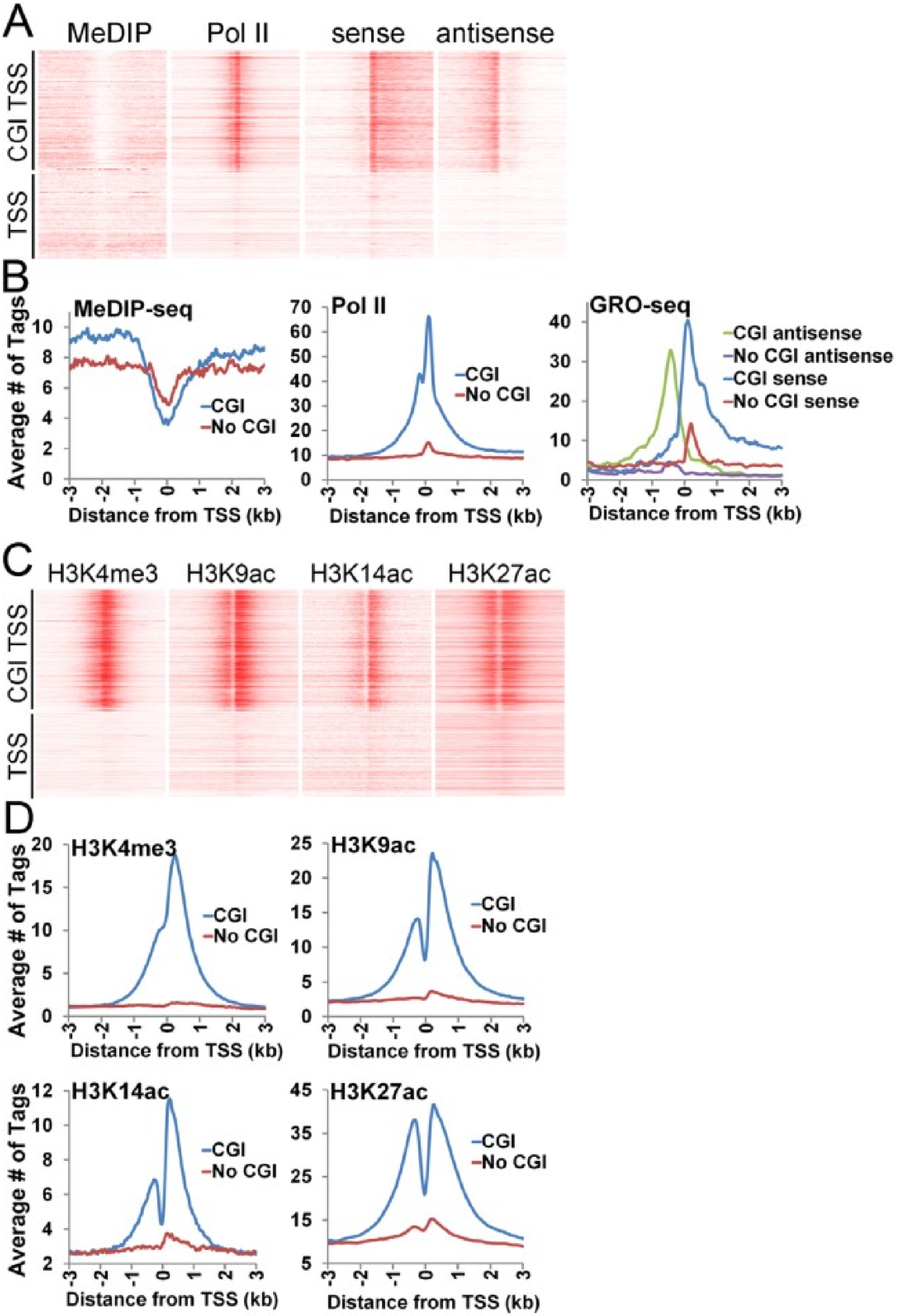

A
MeDIP
Pol II
sense
antisense
CGI TSS
TSS
B
MeDIP-seq
Average # of Tags
0
2
4
6
8
10
-3
-2
-1
0
1
2
3
Distance from TSS (kb)
CGI
No CGI
Pol II
0
10
20
30
40
50
60
70
CGI
No CGI
GRO-seq
0
10
20
30
40
CGI antisense
No CGI antisense
CGI sense
No CGI sense
C
H3K4me3
H3K9ac
H3K14ac
H3K27ac
CGI TSS
TSS
D
H3K4me3
0
5
10
15
20
CGI
No CGI
H3K9ac
0
5
10
15
20
25
CGI
No CGI
H3K14ac
2
4
6
8
10
12
CGI
No CGI
H3K27ac
5
15
25
35
45
CGI
No CGI
Distance from TSS (kb)

Figure S1. Promoter associated CGI are associated with higher levels of divergent transcription.

(A,C) Heat map representation of the tag densities (20bp bins) for DNA methylation (MeDIP-seq, white to red, 0-10), initiated Pol II (phospho–S5- Pol II ChIP-seq, white to red, 0-60), nascent transcription (GRO-seq; sense, antisense, white to red, 0-35); and (C) histone modifications: H3K4me3 (white to red, 0-20), H3K9ac (white to red, 0-25), H3K14ac (white to 3red, 0-10), and H3K27ac (white to red, 0-40) for a 6 kb region surrounding the TSS (+/- 3kb) from MCF7 cells. TSSs were defined as CGI-associated if they were encompassed by an annotated CGI (CGI TSS; n=16,657) or not (No CGI TSS; n= 11,878). (B, D) Average tag densities of +/- 3 kb around for TSSs parsed by absence or presence of CGI from data in (A) and (C) respectively.

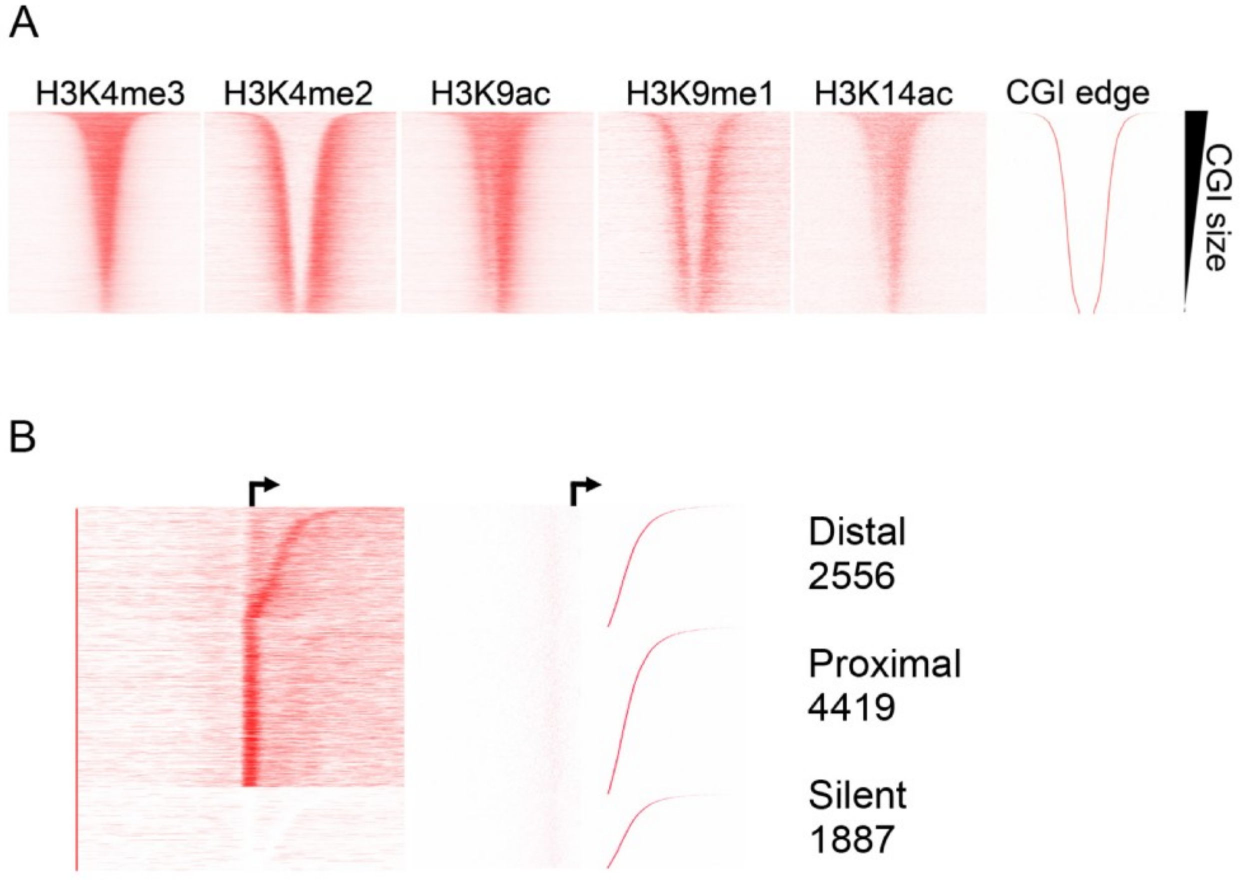


Figure S2. Promoter associated CGI maintain a transcriptionally permissive chromatin domain that is tightly constrained.

Heat map representation of ChIP-seq tag density (20 bp bins) of H3K4me3 (white to red, 0-20), H3K4me2 (white to red, 0-50), H3K9ac (white to red, 0-25), H3K9me1 (white to red, 0-25), and H3K14ac (white to red, 0-10) from MCF7 cells. CGI associated promoters (CGI TSS; n=16,657) were oriented to transcription and sorted by by descending CGI size. Plotted is the density around the midpoint of the CGI +/- 3kb. (B) Heatmap representation of MCF7 GRO-seq sense strand tag density (20 bp bins) across CGI containing a single annotated TSS (white to red, 0- 35). CGI promoters are oriented to transcription and sorted within each pausing class by the distance from the TSS (arrow) to the 3' CGI edge using the same sort order described in Figure 1C.

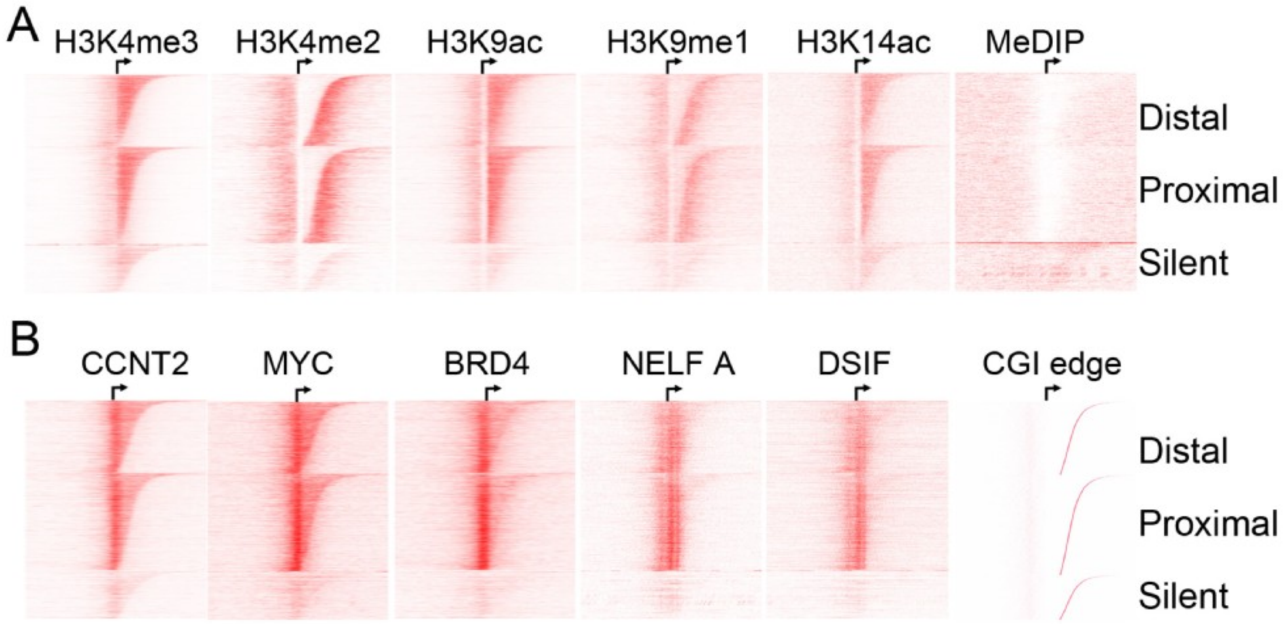


Figure S3. Neither promoter-associated histone modifications nor known pausing factors correlate with Pol II pausing at the distal edge.

(A,B) Heatmap representations of the density (20 bp bins) of (A) histone modifications and DNA methylation and (B) pausing factors for CGI promoters in the three pausing classes. CGI promoters are oriented to transcription and sorted within each class by the distance from the TSS (arrow) to the 3’ CGI edge using the same sort order described in Figure 1C. (A) H3K4me3 (white to red, 0-20), H3K4me2 (white to red, 0-50), H3K9ac (white to red, 0-25), H3K9me1 (white to red, 0-25), H3K14ac (white to red, 0-10) ChIP-seq data and DNA methylation (MeDIP-seq) (white to red, 0-10) data are from MCF7 cells. (B) MYC (white to red, 0-100) and BRD4 (white to red, 0-25) from MCF7 cells; NELFA (white to red, 0-3) and the SUPT5H component of the DSIF complex (white to red, 0-10) from HeLa cells; and CCNT2 (white to red, 0-90) from K562 cells.

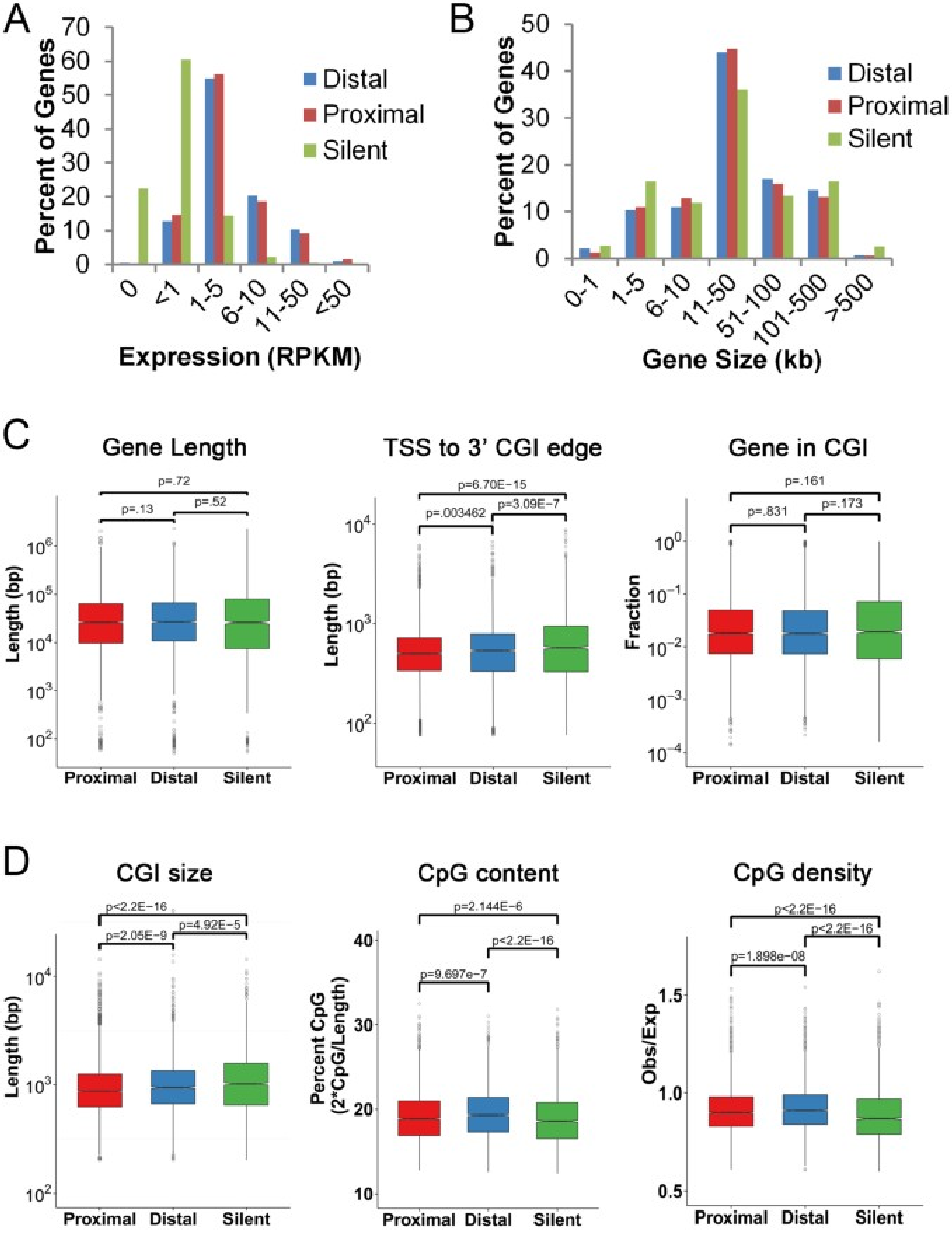

A
Percent of Genes
70
60
50
40
30
20
10
0
0
<1
1-5
6-10
11-50
<50
Expression (RPKM)
Distal
Proximal
Silent
B
Percent of Genes
50
40
30
20
10
0
0-1
1-5
6-10
11-50
51-100
101-500
>500
Gene Size (kb)
Distal
Proximal
Silent
C
Gene Length
p=.72
p=.13
p=.52
Length (bp)
Proximal
Distal
Silent
TSS to 3' CGI edge
p=6.70E-15
p=.003462
p=3.09E-7
Length (bp)
Proximal
Distal
Silent
Gene in CGI
p=.161
p=.831
p=.173
Fraction
Proximal
Distal
Silent
D
CGI size
p<2.2E-16
p=2.05E-9
p=4.92E-5
Length (bp)
Proximal
Distal
Silent
CpG content
p=2.144E-6
p<2.2E-16
p=9.697e-7
Percent CpG (2*CpG/Length)
40
30
20
10
Proximal
Distal
Silent
CpG density
p<2.2E-16
p<2.2E-16
p=1.898e-08
Obs/Exp
1.5
1.0
0.5
Proximal
Distal
Silent

Figure S4. Distally and proximally paused genes do not differ in gene expression levels nor gene characteristics, but show modest differences in CGI features.

(A) Gene expression was measured as RPKM of gene body GRO-seq sense strand tags from MCF7 cells. Shown is the percent of genes in each class falling into the indicated expression levels. (B) Distribution of gene size (distance from the TSS to the TES) is plotted for genes in 7each of the three Pol II pausing classes. (C) Box plots of CGI size, the distance from the TSS to the 3’ CGI edge and the percent of gene covered by the CGI of genes in each pausing class. Median is indicated by a line, box is the 1st and 3rd quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Significance (p-value) was assessed by the Mann-Whitney U test. (D) Box plots of CpG content and CpG density among genes in the three pausing classes. Median is indicated by a line, box is the 1st and 3rd quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Significance (p-value) was assessed by the Mann-Whitney U test.

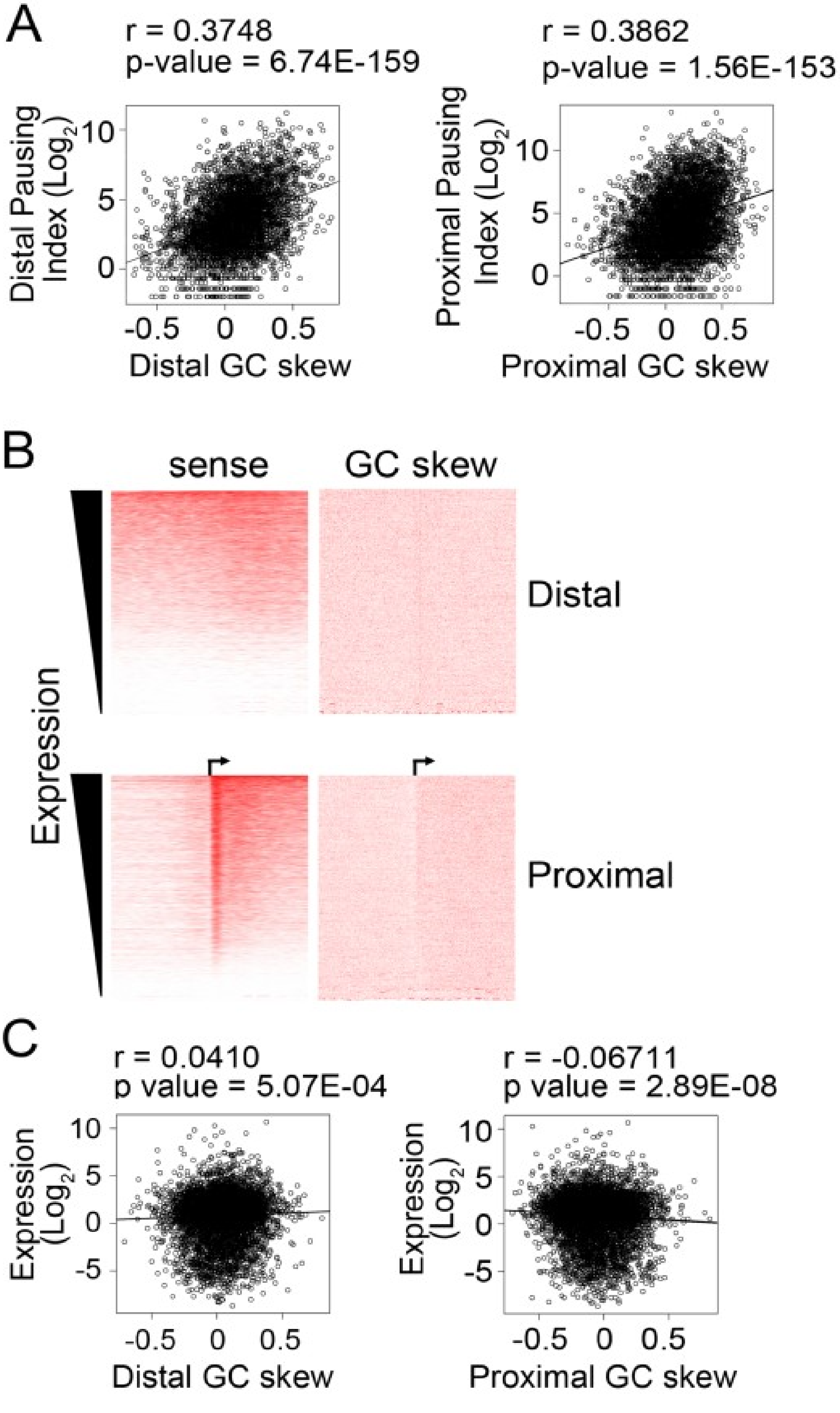

A
r = 0.3748
p-value = 6.74E-159
Distal Pausing Index (Log$_2$)
10
5
0
-0.5
0
0.5
Distal GC skew
r = 0.3862
p-value = 1.56E-153
Proximal Pausing Index (Log$_2$)
10
5
0
-0.5
0
0.5
Proximal GC skew
B
sense
GC skew
Distal
Expression
Proximal
C
r = 0.0410
p value = 5.07E-04
Expression (Log$_2$)
10
5
0
-5
-0.5
0
0.5
Distal GC skew
r = -0.06711
p value = 2.89E-08
Expression (Log$_2$)
10
5
0
-5
-0.5
0
0.5
Proximal GC skew

Figure S5. Pol II pausing correlates better with GC skew than with expression levels. (A) Relationship between GC skew and pausing index at promoter-associated CGI. GC skew was calculated for the 100 bp underlying the distal pause (+20 to +120 bp from the 3' CGI edge, left panel) or the proximal pause (+40 to +140 bp from TSS, right panel) and plotted against log2 of either the distal (left panel) or proximal (right panel) pausing index. Pearsons correlation coefficients are provided. (B) Heat map representation of GRO-seq sense strand tag density (left, white to red, 0-35 tags/20bp) and GC skew (right, white to red, 0-1)) for 6kb centered on the 3' edge of the CGI (top) or the TSS (arrow, bottom). Promoter associated CGI (n= 16,657) were sorted by expression levels as measured by RPKM of GRO-seq sense strand tags across the gene body (3' edge of CGI to TES). (C) Relationship between gene expression and GC skew at the distal (left) or proximal (right) pause. GC skew at the distal (left) or proximal (right) pause was determined for all promoter associated CGI as in A and plotted against log2 of gene expression as measured by the RPKM across the gene body.

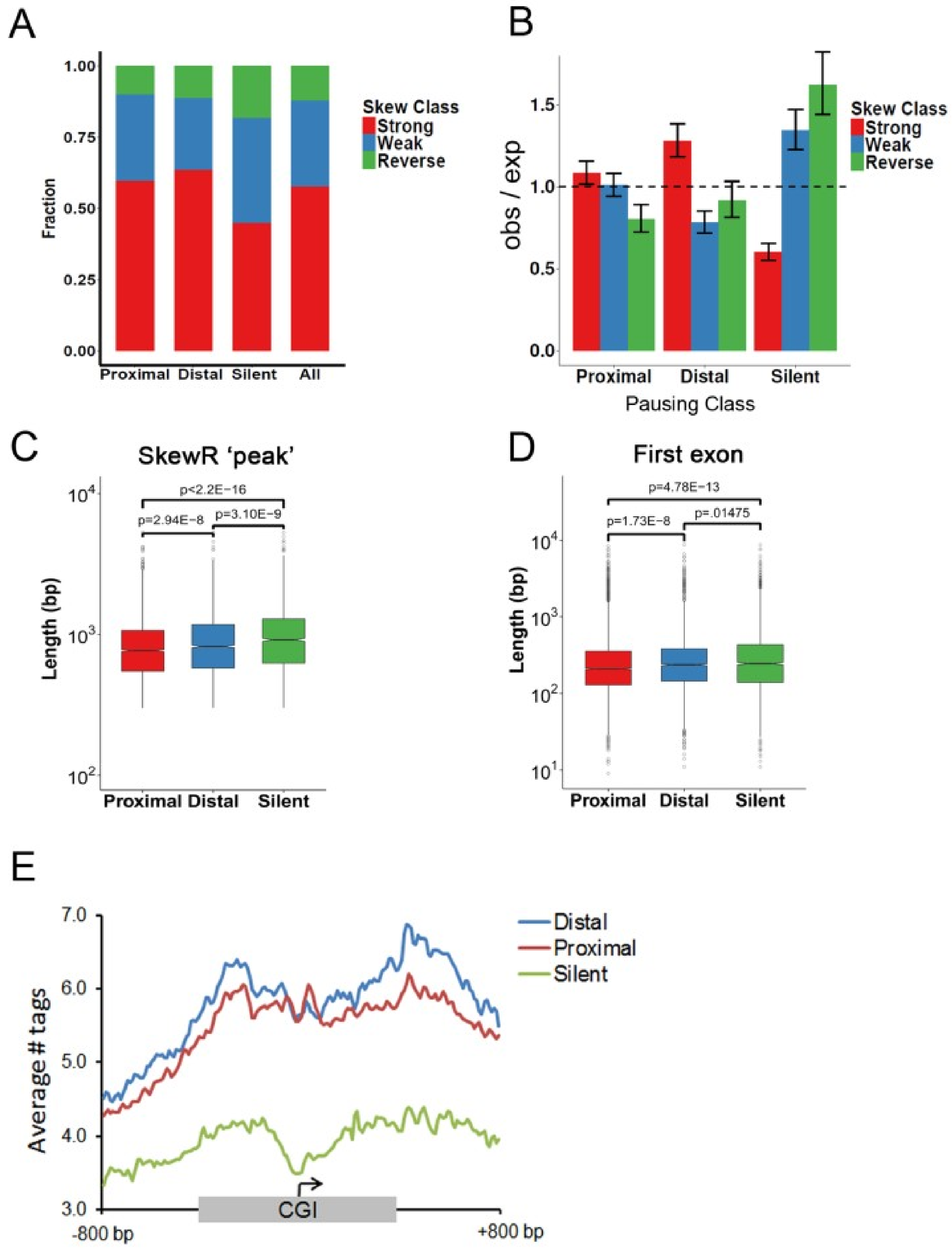

A
1.00
0.75
0.50
0.25
0.00
Fraction
Proximal
Distal
Silent
All
Skew Class
Strong
Weak
Reverse
B
1.5
1.0
0.5
0.0
obs / exp
Proximal
Distal
Silent
Pausing Class
Skew Class
Strong
Weak
Reverse
C
SkewR 'peak'
p<2.2E-16
p=2.94E-8
p=3.10E-9
10^4
10^3
10^2
Length (bp)
Proximal
Distal
Silent
D
First exon
p=4.78E-13
p=1.73E-8
p=.01475
10^4
10^3
10^2
10^1
Length (bp)
Proximal
Distal
Silent
E
7.0
6.0
5.0
4.0
3.0
Average # tags
Distal
Proximal
Silent
CGI
-800 bp
+800 bp

Figure S6: Distally paused CGI genes are associated with longer GC skew regions and are enriched in R-loops downstream of the TSS.

(A) Relationship between pausing class and skew “class” as defined by Ginno et al. (Ginno et al., 2012; Ginno et al., 2013). Shown is the fraction of CGI-associated TSSs in each ‘skew’ class across the 3 pausing classes (Proximal, Distal, Silent) relative to that of unique CGI- associated TSS (All) for which both classes could be called (n=11552) (B) Enrichment of CGI associated TSS with various degrees of skew among the different pausing classes. Bars represent the odds ratio (Fisher’s exact, observed/expected) and whiskers represent the 95% confidence interval. (C,D) Box plots comparing the SkewR peak lengths (C) and length of the first exon (D) among CpG island associated TSSs in each pausing class. CGI associated TSSs were intersected with SkewR ‘peaks’, an HMM that predicts R-loop forming regions based on the degree of GC skew (available from https://www.mcb.ucdavis.edu/faculty-labs/chedin/Resources.html). Approximately 80% of the TSS in each pausing class were found within a SkewR peak (Proximal N=4348, Distal N=2730, Silent N=2641). Median is indicated by a line, box denotes the first and third quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Significance (p-value) was assessed by the Mann-Whitney U test. (E) Average tag densities of RNA:DNA hybrid analysis (DRIP-seq) for promoters in the three pausing classes. CGI associated promoters were oriented to transcription and the distance from the TSS to the upstream and downstream CGI edge independently scaled and anchored to the TSS (arrow). The average tags per 20 bp bin for 800 bp to either side of the CGI (unscaled) is included. Consistent with a lack of transcription, silent genes are depleted for R-loops relative to the other two classes.

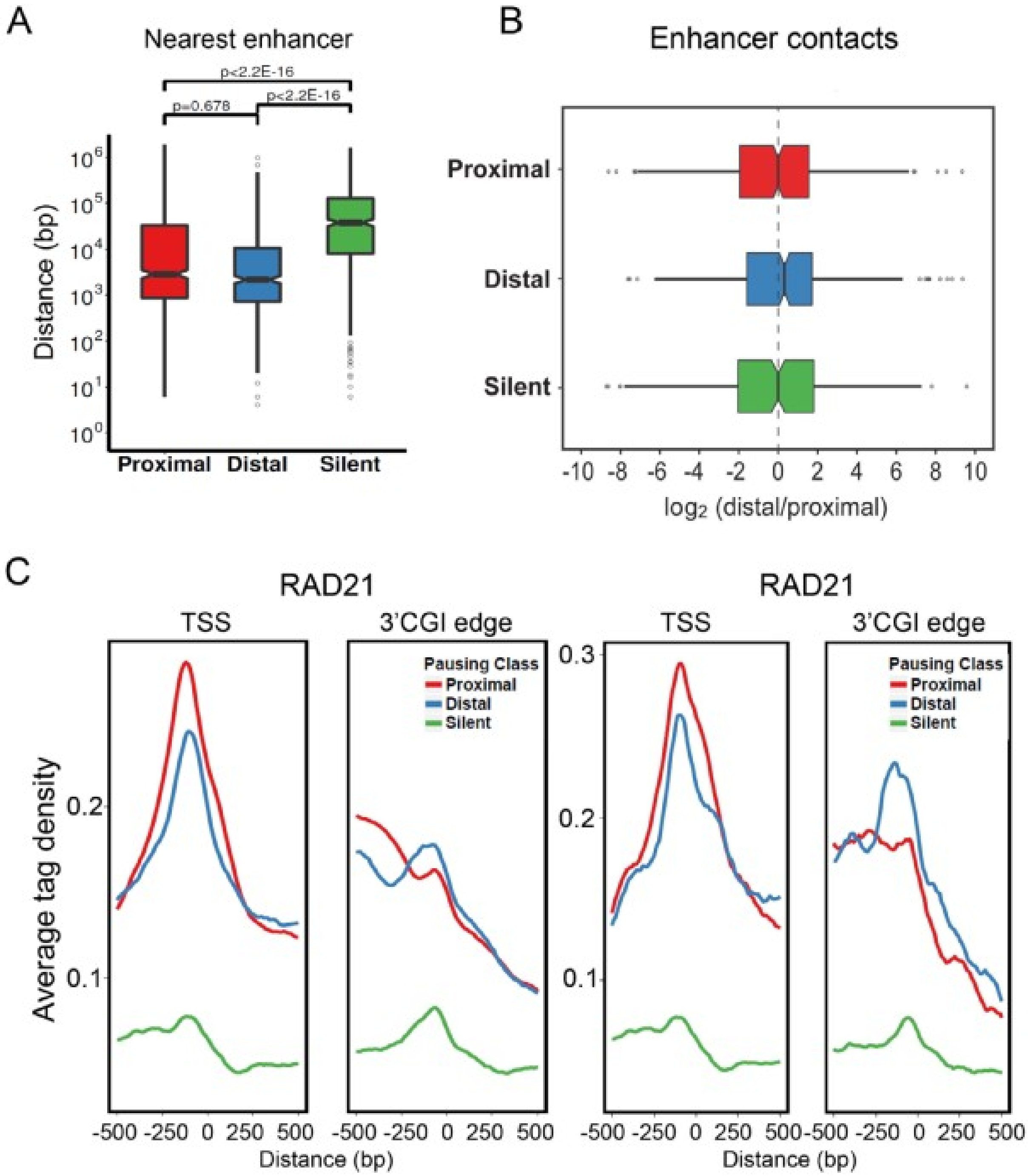

A
Nearest enhancer
p<2.2E-16
p=0.678
p<2.2E-16
Distance (bp)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
Proximal
Distal
Silent
B
Enhancer contacts
Proximal
Distal
Silent
-10
-8
-6
-4
-2
0
2
4
6
8
10
$\log_2$ (distal/proximal)
C
RAD21
TSS
3'CGI edge
Pausing Class
Proximal
Distal
Silent
Average tag density
0.1
0.2
RAD21
TSS
3'CGI edge
0.1
0.2
0.3
Pausing Class
Proximal
Distal
Silent
-500
-250
0
250
500
-500
-250
0
250
500
Distance (bp)
-500
-250
0
250
500
-500
-250
0
250
500
Distance (bp)

Figure S7. Spatial relationship between distal enhancer contacts and pausing class

(A) CGI associated TSSs in each pausing class were annotated to their nearest enhancer, defined as regions of overlapping H3K4me1 and H3K27ac peaks from GM12878 cells. Box plot representation of the distance from the CGI associated TSS to its nearest enhancer for genes in each class. Median is indicated by a line, box denotes the first and third quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. (B) Skewing of the frequency of promoter-enhancer contacts towards the 3'CGI edge in distally paused genes. Hi-C contact matrices at 1kb resolution from GM12878 cells (GSE63525) were used to determine the frequency of contacts between the nearest enhancer and the proximal (TSS+200bp) versus distal (3'CGI edge +/- 100bp) pausing site for each gene. Given the resolution of the Hi-C data, the analysis was limited to those genes for which the TSS and 3'CGI edge are at least 1 kb apart, and that had at least one contact between each pause site and the nearest enhancer (Proximal n=527, Distal n=390, Silent n=308). Shown are box plots of the log2 ratio of enhancer-distal site contacts to enhancer-proximal site contacts among the genes in each pausing class. Median is indicated by a line, box denotes the first and third quartiles, whiskers are the most extreme data point that is 1.5 times the interquartile range. Relative to CGI in the proximal or silent classes, which showed equivalent frequencies contacts at the TSS and distal site, distally-paused genes exhibit a greater frequency of interaction at the distal site (median ratio 1.25 (distally-paused class) vs. 1.0 (proximally paused class), p=.027, Mann-Whitney U test). (C) Average tag densities for the cohesin subunit RAD21 across CGI in the three pausing classes. RAD21 ChIP-seq from MCF7 cells (GSM101079) was used to determine the average tag density per 20 bp bin for +/- 500bp anchored at the TSS or the 3' edge of the CGI, for all promoters in each class (proximal, n=5889; distal n=3663) (left), or the top 20% in each class ranked by pausing index (proximal, n=1289; distal n=738) (right).

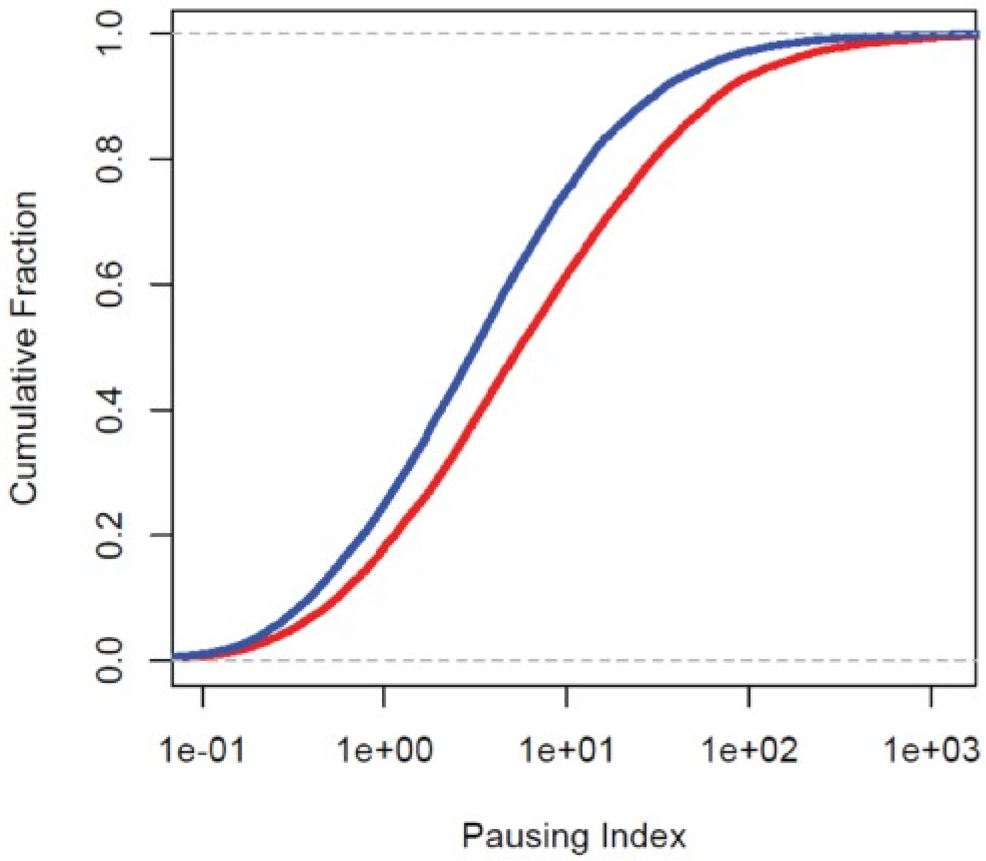


Figure. S8. Relative intensity of the proximal and distal pause.

Cumulative distribution plots showing the pausing index at the proximal (red) or distal (blue) pause points. Pausing index was calculated from GRO-seq tag density in the 100bp underlying the proximal (+40 to +140 bp from TSS) and distal (+20 to +120 bp from the 3' CGI edge) peaks relative to that in the gene body (+120 bp from 3' CGI edge to TES).

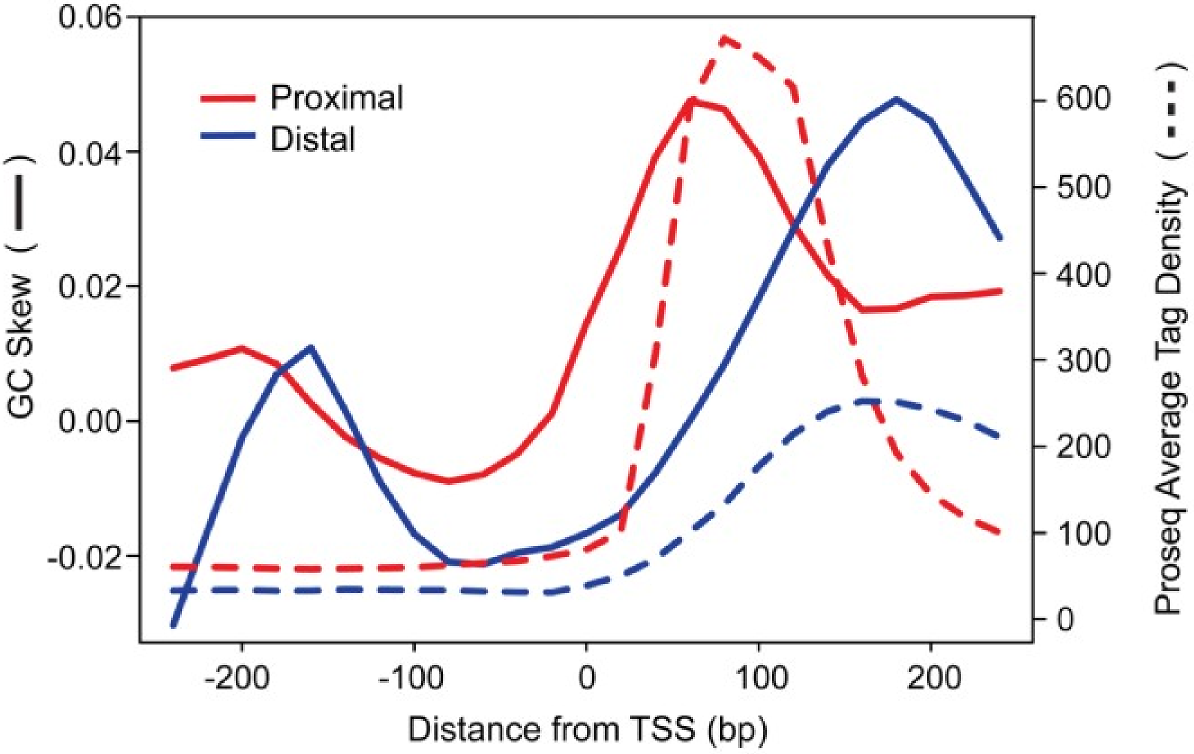
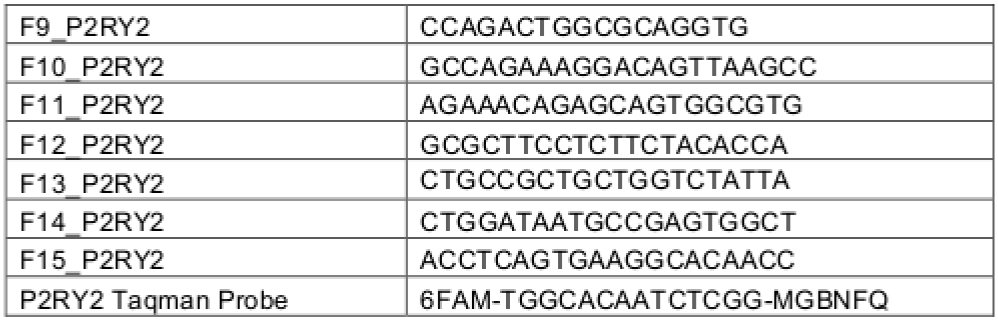


Figure S9. Relationship between GC skew and pausing class in Drosophila

Average GC skew (solid lines) and Pro-seq tag densities (dotted lines) were computed for +/-250 bp surrounding the TSS of Drosophila genes defined as proximally (red) versus distally (blue) paused by Kwak et al. (Kwak et al., 2013). Note the tight relationship between the position of the peak of GC skew and the Pol II pause in each pausing class.

Supplemental Table I – Primers used for 3C Analysis

Gene/Fragment	Sequence
SIAH2_Constant	TCAGATATTAATTGGGTGCCTAGGT
F22_SIAH2	ATAGAAACGACAGCCCTGGGA
F21_SIAH2	CCAACACTTCTTTGGCCCCTA
F20_SIAH2	CTGTGTTCGGAGTCAGGACG
F19_SIAH2	TGCTCCTGAAGGTTTCCACG
F18_SIAH2	CCAAGTTTTACTGGTGCGGC
F17_SIAH2	TTGGTTACACACCAGGTGCC
F16_SIAH2	GTGGTGCTGCGGGGAC
F15_SIAH2	CTTCCTGCTCGGGCTGC
F14_SIAH2	TCAGGACGAGAAGCATTGGG
F7_SIAH2	TGAACCGCATGTCCAAATGT
F1_SIAH2	ATGGCGTAAGAGCCCAGAAG
SIAH2 Taqman Probe	6FAM-CCAGGGACTTCATTG-MGBNFQ
MYC_constant	TGGCTGTTTACCTGGGATCCT
F1_MYC	TGCTCTCTCCTCTGCCGAAA
F4_MYC	GTGACTCACACTGGCAAATTCT
F9_MYC	TATTACCTCCACTACCTGGGGC
F14_MYC	TCCTCCCTGATAGAAGCTCCA
F22_MYC	TACAGCACTTCAAAGCCTCCC
F41_MYC	CATCATCTACAGGGGAGCAGC
F51_MYC	AAACAACCAAGGGTGAGCTACT
F52_MYC	GGGGAAGGGACAACACTAAGC
F53_MYC	TACTGGGCTGGGGTATCAGG
F54_MYC	ACGGAAGTAATACTCCTCTCCTC
F55_MYC	ACTCAGTCTGGGTGGAAGGT
F56_MYC	GACTCTTGATCAAAGCGCGG
F57_MYC	TACTGCGACGAGGAGGAGAA
F58_MYC	CTCCACCTCCAGCTTGTACC
F59_MYC	AGAAATGTCCTGAGCAATCACCT
F60_MYC	ACTTAGAGAGCTCACAGCTTGG
Myc Taqman Probe	6FAMA-CAATGTGTTGCAAGAGT-MGBNFQ
P2RY2_constant	TTGCCCAGGCTGCAATG
F1_P2RY2	CACTGGCCTGGAGATTCAAC
F2_P2RY2	CCTTGGCTGCTTGGTTCCAG
F3_P2RY2	CAGTCAGCTGATATGGAGCCC
F4_P2RY2	CCAGCTCCCTTCTAGCGTG
F5_P2RY2	CAGACACGCTGACCCCG
F6_P2RY2	CTTCGGGGTTGGGGAACAG
F7_P2RY2	GCACCCTGAGAGGAGAAGC



F9_P2RY2	CCAGACTGGCGCAGGTG
F10_P2RY2	GCCAGAAAGGACAGTTAAGCC
F11_P2RY2	AGAAACAGAGCAGTGGCGTG
F12_P2RY2	GCGCTTCCTCTTCTACACCA
F13_P2RY2	CTGCCGCTGCTGGTCTATTA
F14_P2RY2	CTGGATAATGCCGAGTGGCT
F15_P2RY2	ACCTCAGTGAAGGCACAACC
P2RY2 Taqman Probe	6FAM-TGGCACAATCTCGG-MGBNFQ

Factors affecting the Persistence of Drug-induced Reprogramming of the Cancer Methylome

Joshua S.K. Bell[*], Jacob D. Kagey*, Benjamin G. Barwick, Bhakti Dwivedi, Michael T. McCabe, Jeanne Kowalski, and Paula M. Vertino

**These authors contributed equally to this work*

This work is published in *Epigenetics* 2016 Apr 2;11(4).

Contributions: JK performed the experiments. JB analyzed the data and wrote the manuscript.

BG, BD, MM, and JK contributed to the analysis.

Abstract

Aberrant DNA methylation is a critical feature of cancer. Epigenetic therapy seeks to reverse these changes to restore normal gene expression. DNA demethylating agents including 5-aza-2’-deoxycytidine (DAC) are currently used to treat certain leukemias, and can sensitize solid tumors to chemotherapy and immunotherapy. However, it has been difficult to pin the clinical efficacy of these agents to specific demethylation events and the factors that contribute to the durability of response remain largely unknown. Here we examined the genome-wide kinetics of DAC-induced DNA demethylation and subsequent remethylation after drug withdrawal in breast cancer cells. We find that CpGs differ in both their susceptibility to demethylation and propensity for remethylation after drug removal. DAC-induced demethylation was most apparent at CpGs with higher initial methylation levels and further from CpG islands. Once demethylated, such sites exhibited varied remethylation potentials. The most rapidly remethylating CpGs regained >75% of their starting methylation within a month of drug withdrawal. These sites had higher pretreatment methylation levels, were enriched in gene bodies, marked by H3K36me3, and tended to be methylated in normal breast cells. In contrast, a more resistant class of CpG sites failed to regain even 20% of their initial methylation after 3 months. These sites had lower pretreatment methylation levels, were within or near CpG islands, marked by H3K79me2 or H3K4me2/3, and were overrepresented in sites that become aberrantly hypermethylated in breast cancers. Thus, whereas DAC-induced demethylation affects both endogenous and aberrantly methylated sites, tumor-specific hypermethylation is more slowly regained, even as normal methylation promptly recovers. Taken together these data suggest that the durability of DAC response is linked to its selective ability to stably reset at least a portion of the cancer methylome.

Introduction

Although cancer has historically been understood as a disease resulting from genetic mutation, it is now clear that epigenetic alterations are also of critical importance in carcinogenesis. In contrast to genetic mutations, epigenetic changes are potentially reversible, making them provocative therapeutic targets. Among epigenetic therapies, DNA methyltransferase inhibitors in particular have shown activity in the clinic; 5-aza-2'-deoxycytidine (DAC) and 5-azacytidine (AZA) are used to treat myelodysplastic syndrome as well as advanced acute myeloid leukemia and chronic myeloid monocytic leukemia.[1] These drugs have also exhibited activity in solid tumors of the lung in combination with other therapies including HDAC inhibitors [2] and are in clinical trials for the treatment of breast [3] and ovarian cancers. [2, 4-6]

In spite of the considerable use and study of these drugs, the molecular basis of their efficacy remains incompletely understood. As cytidine analogs, DAC and AZA require intracellular conversion to the triphosphate form, and incorporation into DNA where they covalently trap DNA methyltransferase 1 (DNMT1) and lead to its proteasomal degradation.[7] Successive rounds of DNA replication in the absence of DNMT1 results in global hypomethylation. However, the means by which this hypomethylation affects gene expression programs and suppresses tumorigenic potential remain unclear. Although tumor-suppressor genes are frequently silenced in cancer through hypermethylation of promoter-associated CpG Islands (CGI),[8] and DAC is capable of reversing this aberrant methylation to achieve gene reactivation, less than 10% of DAC-induced transcriptional changes are accounted for by relief of promoter hypermethylation. [9, 10]

Cell culture studies have demonstrated that at low doses, transient exposure to DAC results in a loss of tumorigenic potential that persists for many generations following drug withdrawal ,[11] suggesting that durable methylation changes underlie its antitumor activity. Leukemia cells from patients undergoing DAC treatment experience rapid remethylation at the end of each treatment cycle, as measured globally [12] and at LINE elements.[13] Intriguingly, neither the initial genomic methylation level,[14] nor the amount of global demethylation observed upon treatment predicts patient response,[13] suggesting that specific, persistent demethylation events may be responsible for outcome.

Our laboratory and others have documented that while some genes remain stably demethylated after drug withdrawal, others rapidly regain their original methylation state after treatment ends.[10, 15-17] These studies implicated several factors in the propensity of given CpG sites to remethylate including proximity to a transcription start site (TSS), occupancy of RNA Polymerase II (Pol II), and the presence of certain histone modifications.[10, 16] Given the widespread use of these agents and the promising clinical trials underway utilizing DAC or related compounds as a synergistic or sensitizing agents,[4-6] it is vital to understand both the extent and stability of DAC-induced demethylation given its immediate potential to improve cancer treatment.

In this study, we sought to identify the locus-specific determinants underlying the genome-wide susceptibility to DAC-induced demethylation and subsequent remethylation after drug withdrawal. We unearth extensive variation in the sensitivity of CpG sites to demethylation as well as in inclination to remethylation following drug withdrawal. We find that DAC affects more highly methylated CpGs, and that CGI-associated CpGs tend to be resistant to demethylation, in keeping with their generally low methylation levels, but those that do undergo demethylation tend to be slower to remethylate upon drug removal. Moreover, chromatin modifications appear to be predictive of remethylation rate, with chromatin features associated with active promoters, strong enhancers, and elongation rate

(H3K79me2) correlated with resistance to remethylation, while H3K36me3 predisposed associated sites to rapid remethylation. Strikingly, we find that the more methylated a CpG site is in normal breast tissue, the quicker it is to regain that methylation after drug removal, whereas CpGs exhibiting cancer-specific hypermethylation, once demethylated, are slower to recover. These data suggest that DAC treatment stably resets at least a portion of the cancer epigenome to its original state, and substantiates the continued interest in epigenetic therapy as cancer treatment.

Results

Characterization of DAC-induced DNA demethylation

To model the dynamics of DNA demethylation and remethylation in response to DNA methyltransferase inhibitors as it might occur in a solid tumor setting, we treated MDA-MB-231 triple-negative breast cancer cells with 500 nM DAC for six days to induce DNA demethylation, and then tracked DNA remethylation kinetics for over 27,000 CpGs over 27 passages (~3 months) in the absence of drug using Illumina Infinium DNA methylation arrays. DAC-induced demethylation and remethylation kinetics were established for three independent time course experiments. Treatment of MDA-MB-231 cells with DAC elicited a > 25% decrease in genome-wide DNA methylation as estimated from the average β value (Figure 1A). Hierarchical clustering of CpG methylation values segregated untreated and DAC-treated samples by overall methylation levels (Figure 1B, $p< 0.01$ [18]). Examination of the distribution of DNA methylation levels across CpG sites in treated vs. untreated samples revealed that DAC treatment elicited a systematic shift towards lower methylation levels and a decrease in the frequency of highly methylated CpGs ($\beta > 0.7$), complemented with a gain in the frequency moderately (β ~0.4-0.6) and unmethylated sites ($\beta < 0.2$), consistent with global demethylation (Figure 1C).

To identify loci that were differentially methylated after DAC treatment, we used a linear fixed-effects model (see methods). This analysis identified 5,316 loci that were significantly changed upon DAC treatment, all of which were hypomethylated ($FDR \leq 0.05$, $\Delta\beta \geq 0.2$) (Figure 1D,E). Analysis of the distribution of pre- and post-treatment DNA methylation levels for these significantly affected CpGs indicated that DAC tended to affect CpGs that had higher initial levels of DNA methylation, with more than 73% of affected sites having a starting methylation level of >0.7 (Figure 1F). These data demonstrate that DAC induced an overall genome-wide loss in DNA methylation.

Promoters and CpG Islands undergo less demethylation than adjacent genomic regions

Next we determined the spatial relationship between the DAC-affected sites and other genomic features. The position of CpG sites subject to DAC-induced hypomethylation were plotted by their distance to the nearest Transcription Start Site (TSS) and their scaled distance within or to a CGI (Figure 2A,B). Consistent with the tendency for a higher initial DNA methylation level, demethylated CpGs tended to be more distal, and were enriched upstream of the TSS and in the CGI shores as compared to all CpGs represented on the array ($p < 2.2E-16$, Chi-squared). As expected given the relatively unmethylated status of CGI, CpGs undergoing DAC-induced demethylation were underrepresented in CGI.

To determine what chromatin features might influence DAC-induced demethylation, we examined the density of various histone modifications in and around the demethylation-prone CpG sites, utilizing Chip-seq data from human mammary epithelial cells (HMEC) from the ENCODE project.[19] We found that CpGs that underwent significant demethylation were far less likely to be marked by active modifications (H3K4me2/3, H3K9Ac, H3K27Ac, and H2AZ) in normal cells than the other CpG sites (Figure 2C). This finding is consistent with the fact that these marks are enriched near active promoters, which also typically exhibit low levels of DNA methylation. In contrast, CpGs subject to DAC-induced demethylation were more enriched than other analyzed sites in H3K27me3, a repressive mark deposited by the Polycomb complex. This may be a reflection of the propensity for H3K27me3-marked CpGs in normal cells to acquire DNA methylation in cancer cells[20] and thus a high initial methylation level of these sites in general in the MDA-MB-231 cells, or a propensity toward loss of DNA methylation at these sites upon treatment.

Kinetics of CpG remethylation following DAC withdrawal

The above data suggest that both the initial methylation level and the normal underlying chromatin features influence DAC-induced demethylation. We next sought to determine whether, and to what degree, CpG sites differ in their propensity for remethylation after drug withdrawal. To account for differences in initial methylation at each CpG site and amount of methylation lost, we normalized CpGs by their initial and post-treatment DNA methylation levels, such that the degree of demethylation achieved after 6 days of treatment was 100%. DNA methylation changes at later time points (3, 6, 9, and 27 passages) were used to calculate the fractional recovery of methylation over time in the absence of drug. We compared these remethylation kinetics ascertained by the 27k array to methylation measured by COBRA, a restriction enzyme-based method for measuring methylation at individual CpG sites, and found excellent correlation, validating our approach (Supplementary Figure 1). A self-organizing map (SOM) [21] approach was then used to cluster the CpGs into classes based on their patterns of remethylation.

We found that four SOM classes clearly discriminated CpG sites into classes with different remethylation kinetics. We termed these four classes Resistant (N = 1,565), Slow (N

= 1,991), Moderate (N = 1,244), and Rapid (N = 516) based on their remethylation rates (Figure 3A,B). Resistant CpGs represented those relatively resilient to remethylation; most failed to regain even 15% of their lost methylation after 27 passages. Slow CpGs exhibited a sluggish remethylation rate, and on average regained only 30% of their lost methylation, while Moderate CpGs regained ~half of their initial methylation by 9 passages (~1 month). Most striking were Rapid CpGs, which regained more than 70% of their lost methylation within just 9 passages, and almost 90% by 27 passages (~3 months). Comparatively, no other class regained more than 65% of lost methylation even following 27 passages. This analysis demonstrates the considerable variation in remethylation potential and dynamics among CpGs following DAC-induced demethylation.

Recent work by Yang et al.[10] tracked DNA remethylation kinetics following DAC treatment of HCT116 colon cancer cells. An analysis of the remethylation kinetics of the CpGs in our four classes in that experiment showed that these CpGs behave similarly (Figure 3C) in another cell type. These data suggest that the susceptibility of individual CpG sites to demethylation and subsequent remethylation correlates with intrinsic genomic and epigenomic features that are consistent across tissue types.

CpG Islands are resistant to remethylation; shores and gene bodies are prone to remethylation

We next sought to elucidate the factors governing the differences in remethylation rates among CpG sites. First, we addressed the hypothesis that the initial methylation state of CpG sites would correlate with their remethylation rate. Indeed, CpGs in the Rapid class tend to have higher initial methylation levels than Resistant CpGs (Figure 4A, $p < 2.2E-16$, Mann-Whitney U). However, there was no difference in the initial DNA methylation levels among the

Moderate and Rapid loci (mean β =0.82 for each class, p =.55, Mann-Whitney U) suggesting that there are additional factors that govern remethylation rate. Next, we investigated if these methylation differences were linked to differences in proximity to CGI, given the generally low methylation levels of CGI. First, we defined the subset of CpGs in each class associated with genes whose promoters contained CGI (presence of a UCSC-defined CGI within 2 kb of the nearest RefSeq TSS to the CpG site, N = 3,029). Overall, CpGs in CGI-associated genes represented a greater proportion of the Resistant class (67.9%, Odds Ratio (OR)=1.29, p=4.95E-6, Fisher's exact) and there was a trend towards decreasing representation of such sites as remethylation rate increases (Slow, 56.7%, OR=.98, p=.77, Moderate, 50.3%, OR=.85, p=.011, Rapid, 40.9%, OR=.69, p=.00046). We next classified this subset of DAC-demethylated CpG loci by their position within or distance to the CGI. While ~30% of Resistant CpGs were located in a CGI, only half that portion (~15%) of Rapid CpGs were in CGI (OR within CGI: Resistant=1.20, p=.023; Rapid=0.52, p=.0015) (Figure 4B). Indeed, there was a direct relationship between the proximity to the CGI and the remethylation rate. Whereas the Resistant loci tend to be distributed within or immediately downstream of the CGI (the South shore), the Rapid CpGs were found to be more distally distributed (Figure 4C). Slow and Moderate CpGs, intermediate in their remethylation rate, were also intermediate in their distances to CGI (Figure 4C).

To determine whether the differences in remethylation kinetics were driven by the CGI itself or its embedded TSS, we plotted the position of each class of CpGs relative to the nearest TSS. There was no correlation between remethylation rate and distance to the nearest TSS (Figure 4D). Consistent with this finding, there was also no relationship between TSS proximity and remethylation rate among CpGs whose nearest TSS is not CGI-associated (Figure 4E). Together these data suggest that it is characteristics of the CGI domain itself, rather than the embedded TSS, that are responsible for the differences observed in remethylation rate.

Chromatin features correlate strongly with remethylation potential

The above data suggest that one determinant of remethylation potential is the proximity to CGI, which are characterized by a unique chromatin domain.[22] We therefore examined the relationship between remethylation kinetics and local chromatin environment for CpGs in each remethylation class. We first utilized an annotation of chromatin states derived from a Hidden Markov Model (ChromHMM) that uses ChIP-seq data of 9 chromatin marks from HMECs to partition the genome into discrete functional states (Figure 5).[23] We considered both the fraction of sites in each remethylation class that fell into broad categories of chromatin HMM states, as well as the Odds Ratio (OR) of enrichment of CpGs in individual chromatin states. As a second approach, we directly examined ChIP-Seq data for specific histone marks derived from HMECs[19] to investigate the spatial relationship between the enrichment of a given mark around the CpG sites in each remethylation class (Figure 6).

Consistent with the enrichment of Resistant loci in CGI (Figure 4B), Resistant and Slowly remethylating CpGs were far more likely to be associated with ChromHMM states associated with promoter activity in HMECs (Figure 5A, B) and to be enriched in histone marks indicative of active promoters, including H3K4me2/3 and H3K9Ac (Figure 6A), relative to Moderate or Rapidly remethylating loci. In contrast, Moderate and Rapidly remethylating CpGs were depleted in promoter features and display only modest (Moderate) or no (Rapid) enrichment for promoter-associated histone modifications (Figure 6A). As a class, Resistant sites were especially enriched in strong promoters, whereas Slow CpGs were found within many weak and poised promoters (Figure 5C), suggesting that among promoter-associated

CpGs the relative resistance to remethylation may be correlated with promoter activity.

CpGs classified as Rapidly remethylating were strongly enriched in transcribed areas of the genome (Figure 5A), and in particular those marked with features of transcriptional elongation (Figure 5B). Indeed, Rapid CpGs are preferentially enriched in H3K36me3 and H4K20me1 in normal cells (Figure 6B). Rapid CpGs are also notably enriched in the weak transcription/ transcriptional transition ChromHMM category (Figure 5B). Interestingly, the Resistant class is particularly enriched in the gene body mark H3K79me2 (Figure 6B), the levels of which are correlated with Pol II elongation rate.[24, 25]

We next considered the remethylation of CpG sites associated specifically with enhancers, as defined by the ChromHMM classification (Resistant N = 231, Slow N = 366, Moderate N = 209, Rapid N = 75). While a similar fraction of CpGs in each class were annotated to enhancers (Figure 5A), Resistant CpGs that are enhancer-associated are proportionally overrepresented in strong versus weak enhancers, whereas Rapid CpGs displayed the opposite pattern and were proportionally associated with more weak than strong enhancers (Figure 5C). Enhancer activity has been correlated with the extent of local histone acetylation, in particular H3K27Ac levels.[26] Analysis of this mark demonstrated that it is inversely correlated with remethylation rate (Figure 6D) with striking enrichment among Resistant CpGs relative to the other classes. Together, these data suggest that among enhancer-associated CpGs, enhancer activity correlates with resilience against remethylation, mimicking the pattern observed at promoters.

Interestingly, we found that chromatin features associated with the Polycomb complex were also inversely correlated with the propensity for remethylation. The Resistant class exhibited enrichment in Polycomb-silenced regions as defined by the ChromHMM state while the Rapid class was strongly depleted in such regions (Figure 5A,B). This trend was also observed upon analysis of the local enrichment of H3K27me3, the mark deposited by the Polycomb repressive complex, at CpGs of the various classes with greater enrichment of the histone modification in the Resistant class, but little at sites in the Rapid class. (Figure 6C). Gene Ontology analysis also indicated that the Resistant class was highly enriched in polycomb targets (Supplementary Table 1),   We also find that enrichment in H3K9me3, a prominent marker of heterochromatin and gene repression, correlates with resistance to remethylation when considered independently (Figure 6C).

DAC induces stable reversal of cancer-specific hypermethylation

Previous work has shown that low dose DAC treatment of cancer cells leads to a durable loss of tumorigenic potential for many passages following treatment even when global methylation levels appear largely restored.[11] This suggests that there may be some component of the cancer methylome whose demethylation is preferentially reset to a more normal pattern. At the same time, one concern is the potential impact on normal methylation patterns. To address these questions, we defined a set of normally methylated CpGs as those highly methylated in HMEC cells (β > 0.7; N = 2,326) (Resistant N = 429, Slow N =794, Moderate N = 711, Rapid N = 392). We then determined the relative enrichment for these sites in each remethylation class (Figure 7A). We found that these normally methylated sites were much more likely to Rapidly remethylate compared to other sites, indicating that where endogenous methylation is lost during DAC treatment, it is quickly regained. We further defined a set of CpG sites that undergo cancer-specific hypermethylation as those that were hypermethylated in MDA-MB-231 cells relative to HMEC cells (change in β ≥ 0.2, MDA-MB-231 cells versus HMEC). This identified 2,464 CpG sites (Resistant N = 944, Slow N = 1,011, Moderate N = 428, Rapid N = 81). Strikingly, the Resistant class was heavily enriched in such sites, whereas the Rapid class was markedly depleted, suggesting that methylation gained during tumorigenesis is less likely to be regained after DAC-induced demethylation (Figure 7B).

We next examined the methylation status of CpG sites in each remethylation class among a collection of 90 matched primary breast tumor-normal pairs (N = 180 samples) for whom DNA methylation data was collected as part of the TCGA project (35). We limited our analysis to the subset of our significantly demethylated CpGs common between the platform used here (Illumina 27K array) and that used for the TCGA samples (Illumina 450K array). This resulted in a total of 3,989 CpG sites (Resistant = 1,213, Slow = 1,527, Moderate = 906, Rapid = 342). An unsupervised hierarchical clustering approach showed that the methylation status of CpG sites in each class was independently capable of segregating tumor from normal samples (Resistant, $p < 1.25x10^{-37}$; Slow, $p < 8.22x10^{-37}$; Moderate $p < 1.82x10^{-38}$; Rapid, $p < 1.25x10^{-37}$) and each outperformed 1,000 randomly-selected sets of the same size in this regard ($p < 0.0001$) indicating that the CpG sites in each remethylation class capture at least a portion of the cancer-specific methylome (Figure 7C). Intriguingly, an examination of the distribution of DNA methylation levels among CpGs in each class in normal breast tissue samples showed that there was a significant difference in the normal methylation levels among CpGs in the different remethylation classes, with the Rapidly remethylating CpGs having twice the normal DNA methylation levels of Resistant loci (Resistant: median=0.41, Rapid: median=0.85, p<2.2E-16, Mann-Whitney U) (Figure 7D). The relationship between remethylation rate and normal tissue methylation levels was even more striking than that observed in untreated MDA-MB-231 cells (compare Figure 4A and Figure 7D). These data confirm the above finding that the more methylated a CpG site is in normal tissue, the more likely it is to quickly regain this methylation after DAC-induced demethylation, and further suggest that the methylation state of a CpG in normal tissue is more predictive of its resilience to remethylation than its pretreatment methylation state in cancer cells.

To more directly address the relationship between cancer-specific hypermethylation and remethylation potential, we defined a set of CpGs that were significantly hypermethylated in the breast tumors relative to the matched normal tissues using a linear mixed-effects model approach.[27] Among the 25,978 CpG sites common to the two analytic platforms (Illumina 27K vs. 450K), a total of 2,356 sites were identified as significantly hypermethylated using this approach, including 440 Resistant, 595 Slow, 328 Moderate and 98 Rapid. Interestingly, we found that CpGs hypermethylated in cancer were depleted in Rapidly remethylating CpGs relative to other classes (Figure 7E). Together, these data suggest that cancer-specific hypermethylation is more likely to be regained at a slower rate than normal methylation, perhaps contributing to the lasting effect of demethylating agents.

Discussion

DNA methyltransferase inhibitors are currently the standard of care for certain myeloid malignancies and are showing promise alone and in combination with other therapies in the treatment of solid tumors.[28, 29] A number of mechanisms have been proposed to account for the antitumor activity of DNA methyltransferase inhibitors, including the reactivation of cancer testes antigens stimulating an immune response,[30] reactivation of silenced tumor suppressor genes, repression of oncogenes through loss of gene body methylation,[10] and most recently, the activation endogenous retroviruses, leading to the cytoplasmic accumulation of double-stranded RNAs and the triggering of an antiviral interferon response.[31, 32] In solid tumor model systems low-dose DAC treatment has been shown to result in prolonged demethylation and a sustained, heritable inhibition of the tumorigenic potential of cancer-initiating/stem-like populations. [10, 11, 32] Thus, it is vital to understand the site-specific and global factors influencing DNA demethylation and remethylation kinetics as this may provide insight into the mechanisms of drug action and potentially differential responses, which have thus far been difficult to predict.[29]

Our data indicate that CpGs across the genome differ in both their susceptibility to DAC-induced demethylation as well as their propensity for remethylation after drug removal. We find that CpG islands are not the major targets of DAC-induced demethylation, in line with their low initial methylation, and consistent with the observation that few transcriptional changes in DAC treatment are explained by relief of promoter CGI hypermethylation.[9] Rather, we find that upstream shore and shelf regions are much more likely to undergo demethylation than CGI. Recent work has shown that much of the cancer-specific variation in DNA methylation patterns occurs in these shore regions.[33] Reversal of cancer-specific methylation in these areas in particular may thus play an important role in determining the therapeutic activity.

Chromatin states appear to have profound influence on both DAC-induced demethylation and remethylation rate. Regions that are marked by active chromatin marks (eg. H3K4me2/3, H3K9/27Ac, and H2AZ) in normal mammary epithelial cells, such as CGI, tend to be resistant to demethylation, and those that do undergo demethylation are resistant to remethylation upon drug removal. Indeed, remethylation-prone CpGs tend to be excluded from CGI. Whether it is transcription itself, RNA Pol II occupancy, or the chromatin architecture of the CGI that is the primary deterrent is difficult to uncouple, but it is noteworthy that remethylation rate appears to be more tightly linked to the proximity to or presence within the CGI domain than to the TSSs within those CGI, suggesting that it is the CGI structure rather than transcription *per se* that plays a key role in determining propensity to remethylation. Indeed, CpG site proximity to the TSS of non-CGI associated genes had little impact on remethylation rate. CGIs represent unique chromatin environment in the genome, and are characterized by constrained divergent (bidirectional) transcription, marking by H3K4me3, and high levels of GC-skew, a sequence-based feature associated with the formation of unusual secondary structures.[34] Transcription through such regions leads to the formation of R-loops formed by the pairing of a G-rich nascent RNA to its C-rich template behind the progressing polymerase. R-loops have been suggested to play a key role in preventing DNA methylation at CGI.[35] Our group has recently shown that GC skew additionally defines key points of RNA Pol II pausing in CGI promoters,[36] regulating its release into elongation. Even in its paused state, Pol II occupancy is sufficient to impede remethylation after DAC-induced demethylation.[16]

A similar trend was observed at enhancers; CpGs associated with strong enhancers and high levels of H3K27Ac in normal cells were found to be more resistant to remethylation. Like CGI, active enhancers are associated with bidirectional transcription, though of non-coding enhancer RNAs (eRNA), the levels of which are correlated with enhancer activity.[37] Transcription or Pol II occupancy in these regions may also contribute to remethylation resistance. Interestingly, a significant proportion of the gene expression changes induced by severe hypomethylation in DNA methyltransferase-deficient cells has been linked not to the demethylation of promoters, but rather to the activation of intergenic enhancer regions.[38] The resistance of such demethylated and hyperacetylated loci to remethylation could in part contribute to the synergistic activation of gene expression observed with the combination of DNA methyltransferase inhibitors and HDAC inhibitors and perhaps the treatment benefit observed in some clinical settings.[39-41]

We find that CpGs embedded in regions of H3K27me3 in normal cells are more sensitive to demethylation, yet resistant to remethylation. While there is a great deal of literature suggesting that Polycomb occupancy predisposes to DNA methylation in a developmental context [42] and to aberrant hypermethylation in cancer cells,[20,43] the two marks are largely mutually exclusive genome-wide in differentiated cells, especially in CGI.[44, 45] Indeed, recent work has demonstrated that genomic regions that lose DNA methylation after azacytidine treatment or deletion of *Dnmt1*, tend to gain H3K27me3.[46] This implies a model in which loss of DNA methylation enables restoration of PRC2 occupancy and deposition of H3K27me3, which protects from DNA remethylation. A similar phenomenon may be occurring with H3K9me3. While studies have indicated that H3K9me2/3 is required for the maintenance of DNA methylation [43] and participates in the direction of DNMT1 during replication,[28, 47] recent work has demonstrated that H3K9me3 and DNA methylation generally repress distinct sets of genes.[48] That DAC-induced loss of H3K9me3 from repressed areas often leads to its replacement by H3K27me3 [9] suggests that the relationship between H3K9me3 and remethylation rate may be indirect, and a function of the accumulation of H3K27me3 in hypomethylated regions.

Among the most striking findings was the relationship between CpG remethylation rate and chromatin features associated with gene bodies and transcribed sequences. We find that H3K36me3 correlates well with the rate of remethylation and that the rapidly remethylating class is strongly enriched in chromatin features associated with transcriptional elongation. One possible mechanism for this is the recognition of H3K36me3 by the PWWP domain of the de novo DNA methyltransferases DNMT3A or DNMT3B [49]. Recent work suggests that DNMT3B1 in particular is selectively targeted to the bodies of transcribed genes in an H3K36me3 and PWWP dependent manner. [20] In this way, H3K36me3 may serve as a redundant memory of the lost DNA methylation, allowing rapid remethylation of these genomic regions following drug removal. This suggests that inhibition of SETD2 [50], the methyltransferase responsible for H3K36me3,[51] or blockade of the DNMT3A PWWP domain may synergize with DAC in cancer cell reprogramming. Interestingly, while H3K36me3 and the density of DNA methylation in gene bodies correlates with steady-state transcript levels, they do not appear to be related (H3K36me3) or are negatively correlated (DNA methylation) with the rate of Pol II elongation within gene bodies. [24,25] In contrast, we find that H3K79me2, which has been directly linked to Pol II elongation rate [25] correlates with resistance to remethylation. Together, these data support the hypothesis that the function of DNA methylation and H3K36me3 in gene bodies is to provide a heritable memory of prior transcription and to suppress Pol II initiation at cryptic start sites,[52] rather than to play a role in Pol II elongation.

We also performed gene ontology (GO) analysis to in order to address possible disparate functions of genes in each remethylation category. The most striking and statistically significant finding from this analysis was the enrichment of H3K27me3 targets in the Resistant class, mirroring the enrichment seen in the chromHMM Polycomb state and ChipSeq analyses (Figs. 5 &6). The lack of strong evidence for association with genes of any specific function, however, suggests that the remethylation kinetics of each CpG site are more linked to local chromatin environment and proximity to genetic features, as we detail in this study, than to gene function itself.

It has been proposed that cancer cells become addicted to at least some oncogenic driver methylation events acquired during tumorigenesis, and that the demethylation of such sites might contribute to the anticancer activity of DAC. [29] Critically, we find that the DAC-induced demethylation of CpGs that undergo cancer-specific hypermethylation, whether in the breast cancer cell line or in primary tumors, have a tendency to be more resilient against remethylation, whereas sites that are heavily methylated in normal cells tend to rapidly recover their lost methylation. This implies that remethylation is directed specifically towards those CpGs that are normally methylated, while the aberrant methylation acquired in cancer cells is slower to recover or largely forgotten. The propensity for such “slow” sites to remain demethyated may contribute to the antitumor effects of demethylating agents. The swift remethylation of endogenous methylation may ensure that DAC treatment does not permanently disrupt the normal epigenome, and may explain in part why cancer cells are more sensitive to demethylating agents.[11] Consistent with this idea, CpGs found to be strongly resistant to demethylation and to be required for cell survival in a genetic model of DNMT1/3b depletion [29] also tended to be resistant to DAC-induced demethylation in our study, and the few that were demethylated, are strongly enriched in the rapidly remethylating class (OR 7.5, p=4E-16, data not shown).

In sum, these finding suggest that similar genomic factors may govern sensitivity to programmed *de novo* methylation in development and remethylation in DAC-treated cancer cells. While promoters are rarely methylated in normal cells, gene bodies are dynamically and heavily methylated throughout development, and CpGs that specifically gain aberrant methylation in cancer cells are unlikely to regain that methylation following DAC treatment, even while normal methylation is immediately replaced. This restoration of the endogenous methylome is the express goal of epigenetic therapy, and thus our work reinforces the idea of epigenetic reprogramming and sustained interest in the application of demethylating agents in cancer treatment.

Methods

Cell culture and DAC treatment

Human mammary epithelial cells (HMECs) were obtained from Clonetics and maintained in Mammary Epithelial Cell Growth Medium (Lonza). MDA-MB-231 breast cancer cells were obtained from ATCC. Cells ($4.5\text{x}10^5$) were plated in 10 cm plates and treated the following day with 500 nM DAC every other day for six days.[16] Cells were then maintained in the absence of DAC for 27 passages, with cells being passed 1:10 every three days. DNA was harvested with the Qiagen DNA extraction kit before treatment, immediately following treatment, and at 3, 6, 9, and 27 passages in drug-free media. Three independent biological time-course experiments were performed.

COmbined Bisulfite Restriction Analysis(COBRA)

Bisulfite-modified DNA was amplified using primers devoid of any CpGs. Amplified products were purified with the PCR Purification kit (Qiagen), digested overnight at 37 °C with *FNU4*HI or *Xmn*I, precipitated, and resolved on a 2.0% agarose gel ([23]). Relative intensities of digested and undigested bands were quantified with Image Quant 5.2 and percent methylation was determined as the combined intensity of the digested bands relative to that of all bands (undigested and digested). Primer sequences are in Supplemental Table 2.

DNA methylation analysis

Genomic DNA was bisulfite converted and hybridized to the Illumina Infinium HumanMethylation27 BeadChip by the Emory Integrated Genomics Shared Resource. β values were exported from Genome Studio and analyzed in R / Bioconductor.[53] Loci differentially methylated in response to DAC treatment were determined using a linear fixed-effects model as previously described. [54] Briefly, the "lm" function of the stats package in R / Bioconductor was used to determine the significance of DNA methylation changes. For clustering of 27K arrays, the Pvclust R package was used [18] with the average agglomerative method of hierarchical clustering and correlation distance measure with 10000 bootstrap replications. Pvclust calculates the approximately unbiased (AU) p-value via multiscale bootstrap resampling and the bootstrap probability (BP) p-value based on normal bootstrap resampling. Both were < 0.01 in parsing untreated from DAC-treated methylation arrays. *P*-values were corrected for multiple hypothesis testing using Benjamini-Hochberg False Discovery Rate (FDR) correction.[55] Additionally, a minimum change in DNA methylation of $\beta \geq 0.2$ was imposed. Data was analyzed using the “heatmap.2” function of the gplots package of R / Bioconductor. Genome-wide (tag) density plots and boxplots were generated using the ggplot2 package in R.

Self-organizing maps

To define patterns of re-methylation kinetics, the raw methylation data (β values) from each of the three time course experiments was first combined to obtain the average ㄴㅈ value. The data for each CpG was then normalized to both its starting methylation level as extrapolated from the β values and the degree of demethylation by setting the initial β value for each probe to 1, the post-treatment β value to 0, and determining the fraction of lost methylation recovered at each passage after drug removal. Self-organizing maps were defined using GenePattern [56] with the following parameters: seed range = 42, 100,000 iterations, random vectors initialization, Gaussian neighborhood function, $\alpha_{initial} = 0.1$, $\alpha_{final} = 0.005$, $\sigma_{initial} = 5.0$, $\sigma_{initial} = 0.5$.

Analysis of DNA methylation in primary breast tumor-normal pairs

Level 3 Illumina Infinium HumanMethylation 450K BeadChip methylation data for the four SOM classes: Resistant, Slow, Moderate, and Rapid was extracted for 90 matched tumor-normal pairs of breast carcinoma (BRCA) from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). The methylation data are preprocessed and normalized ㄴㅈ ㄷvalues are available from TCGA. In total, 3,988 CpG sites were considered across the SOM classes that were present on both the Illumina 450K and 27K array. The analysis was restricted to CpG sites with a detection p-value > 0.05 across the dataset (n=386,512), which included Resistant (n = 1,202), Slow (n = 1,517), Moderate (n = 901), and Rapid (n = 336). Unsupervised hierarchical clustering was performed using the Pearson correlation distance with agglomerative complete linkage for CpG sites in each SOM class across all 180 samples. We assessed the statistical significance of the SOM-defined CpG sites in separating breast tumors from normal versus a randomly-defined 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 remethylation class from among the 386,512 CpG sites remaining after filtering for low quality probes. For each re-sampling, dendograms were constructed 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 defined by comparing the p-value obtained from an association analysis of tumor and normal samples to the distribution of p-values obtained based on the two clusters formed using randomly-sampled CpG sites.

CpG sites hypermethylated in primary breast tumors relative to normal tissue were identified from among the probes present on both arrays (N=25,978) using a linear mixed effects model as implemented in CpGAssoc [27] with an FDR cutoff of 0.05. This analysis resulted in 2,356 hypermethylated sites, of which 440 were Resistant, 595 Slow, 328 Moderate and 98 Rapid. The remaining 895 were not among those significantly demethylated in our study. The odds ratio of enrichment of normally methylated CpGs, or those hypermethylated in breast cancers, among the four remethylation classes was determined by

Fisher's exact. CpG sites with a β value > 0.7 in HMEC cells were considered to represent normally methylated sites (Total = 2,326, Resistant N = 429, Slow N = 794, Moderate N =711, Rapid N = 392). Sites were considered to have undergone cancer-specific hypermethylation if the average β value in untreated MDA-MB-231 cells was greater than that observed in HMEC by at least 0.2 (2,464 Total CpG sites, Resistant N = 944, Slow N = 1,011, Moderate N = 428, Rapid N = 81)

Chromatin analyses

HMEC ChIP-Seq data sets were obtained from the ENCODE project. Accession numbers are GSM646374 (H3K27Ac), GSM646376 (H3K27me3), GSM646378 (H3K36me3), GSM646380 (H3K4me1), GSM646382 (H2K4me2), GSM646384 (H3K4me3), GSM646386 (H3K9Ac), and GSM646388 (H4K20me1). Tag densities for each ChIP-Seq were calculated in 20 bp bins for the 5,000 bp surrounding the CpG sites of interest using the GenomicRanges R package [57], and normalized to total read count in each data set.

ChromHMM annotations were based on the classifications reported by Ernst et al. [23] for HMECs and were downloaded from the UCSC genome browser. For some analyses ChromHMM classes of similar function were collapsed; "Promoter" was defined as ChromHMM classes 1-3, "Enhancer" as classes 4-7, "Heterochromatin" as classes 8 & 13-15, "Transcribed" as classes 9-11, and "Polycomb" as class 12. For the odds ratios calculations, the "Strong Enhancer" (4 &5) and "Weak Enhancer" (6 & 7) categories were merged, as were the "Transcriptional Elongation" (9) and "Transcriptional Transition" (10) categories. No CpGs in the list of significantly demethylated sites were found in the Repetitive/CNV ChromHMM categories.

Acknowledgements

The authors wish to thank Michael Nichols for helpful discussions. Supported by NIH grants R01-CA077337, R01-CA132065 (to PMV); NIH NRSA predoctoral fellowships F31-CA186676 (to JSKB) and F31-AI112261 (to BGB), a Department of Defense/CDMRP Breast cancer predoctoral fellowship CDMRP BC073543 (to JDK), and an American Cancer Society postdoctoral fellowship PF-07-130-01-MGO (to MTM). This research project was supported in part by the Winship Biostatistics and Bioinformatics Shared Resource of Emory University and the Winship Cancer Center Support grant P30 CA0138292.

References

1. Navada SC, Steinmann J, Lübbert M, Silverman LR. Clinical development of demethylating agents in hematology. J. Clin Inv 2014; 124:40-6.
2. Juergens RA, Wrangle J, Vendetti FP, Murphy SC, Zhao M, Coleman B, Sebree R, Rodgers K, Hooker CM, Franco N. Combination epigenetic therapy has efficacy in patients with refractory advanced non–small cell lung cancer. Cancer Disc 2011; 1:598-607.
3. Katz TA, Huang Y, Davidson NE, Jankowitz RC. Epigenetic reprogramming in breast cancer: from new targets to new therapies. Ann Med 2014; 46:397-408.
4. Fu S, Hu W, Iyer R, Kavanagh JJ, Coleman RL, Levenback CF, Sood AK, Wolf JK, Gershenson DM, Markman M. Phase 1b-2a study to reverse platinum resistance through use of a hypomethylating agent, azacitidine, in patients with platinum-resistant or platinum-refractory epithelial ovarian cancer. Cancer 2011; 117:1661-9.
5. Matei D, Fang F, Shen C, Schilder J, Arnold A, Zeng Y, Berry WA, Huang T, Nephew KP. Epigenetic resensitization to platinum in ovarian cancer. Cancer Res 2012; 72:2197-205.

6. Wang Y, Cardenas H, Fang F, Condello S, Taverna P, Segar M, Liu Y, Nephew KP, Matei D. Epigenetic targeting of ovarian cancer stem cells. Cancer Res 2014; 74:4922-36.

7. Patel K, Dickson J, Din S, Macleod K, Jodrell D, Ramsahoye B. Targeting of 5-aza-2′-deoxycytidine residues by chromatin-associated DNMT1 induces proteasomal degradation of the free enzyme. Nucl Acids Res 2010; 38:4313-24.

8. McCabe MT, Brandes JC, Vertino PM. Cancer DNA methylation: molecular mechanisms and clinical implications. Clin Cancer Res 2009; 15:3927-37.

9. Komashko VM, Farnham PJ. 5-azacytidine treatment reorganizes genomic histone modification patterns. Epigenetics 2010; 5:229-40.

10. 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.

11. Tsai H-C, Li H, Van Neste L, Cai Y, Robert C, Rassool FV, Shin JJ, Harbom KM, Beaty R, Pappou E. Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. Cancer Cell 2012; 21:430-46.

12. Stresemann C, Bokelmann I, Mahlknecht U, Lyko F. Azacytidine causes complex DNA methylation responses in myeloid leukemia. Mol Cancer Ther 2008; 7:2998-3005.

13. Fandy TE, Herman JG, Kerns P, Jiemjit A, Sugar EA, Choi S-H, Yang AS, Aucott T, Dauses T, Odchimar-Reissig R. Early epigenetic changes and DNA damage do not predict clinical response in an overlapping schedule of 5-azacytidine and entinostat in patients with myeloid malignancies. Blood 2009; 114:2764-73.

14. Shen L, Kantarjian H, Guo Y, Lin E, Shan J, Huang X, Berry D, Ahmed S, Zhu W,

Pierce S. DNA methylation predicts survival and response to therapy in patients with myelodysplastic syndromes. J Clin Oncol 2010; 28:605-13.

15. Bender CM, Gonzalgo ML, Gonzales FA, Nguyen CT, Robertson KD, Jones PA. Roles of cell division and gene transcription in the methylation of CpG islands. Mol Cell Biol 1999; 19:6690-8.

16. Kagey JD, Kapoor-Vazirani P, McCabe MT, Powell DR, Vertino PM. Long-term stability of demethylation after transient exposure to 5-aza-2′-deoxycytidine correlates with sustained RNA polymerase II occupancy. Mol Cancer Res 2010; 8:1048-59.

17. Lin JC, Jeong S, Liang G, Takai D, Fatemi M, Tsai YC, Egger G, Gal-Yam EN, Jones PA. Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. Cancer Cell 2007; 12:432-44.

18. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 2006; 22:1540-2.

19. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012; 489:57-74.

20. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nature Genet 2007; 39:232-6.

21. Kohonen T. The self-organizing map. Proceedings of the IEEE 1990; 78:1464-80.

22. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Dev 2011; 25:1010-22.

23. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M. Mapping and analysis of chromatin state dynamics in nine

human cell types. Nature 2011; 473:43-9.

24. Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife 2014; 3:e02407.

25. Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, Ljungman M. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res 2014; 24:896-905.

26. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009; 459:108-12.

27. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinformatics 2012; 28:1280-1.

28. Ahuja N, Easwaran H, Baylin SB. Harnessing the potential of epigenetic therapy to target solid tumors. J Clin Inv 2014; 124:56-63.

29. Treppendahl MB, Kristensen LS, Grønbæk K. Predicting response to epigenetic therapy. J Clin Inv 2014; 124:47-55.

30. Odunsi K, Matsuzaki J, James SR, Mhawech-Fauceglia P, Tsuji T, Miller A, Zhang W, Akers SN, Griffiths EA, Miliotto A, et al. Epigenetic potentiation of NY-ESO-1 vaccine therapy in human ovarian cancer. Cancer Immunol Res 2014; 2:37-49.

31. Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, Hein A, Rote NS, Cope LM, Snyder A. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. Cell 2015; 162:974-86.

32. Roulois D, Yau HL, Singhania R, Wang Y, Danesh A, Shen SY, Han H, Liang G, Jones PA, Pugh TJ. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. Cell 2015; 162:961-73.

33. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nature Genet 2009; 41:178-86.

34. Ginno PA, Lim YW, Lott PL, Korf I, Chédin F. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res 2013; 23:1590-600.

35. Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 2012; 45:814-25.

36. Kellner WA, Bell JS, Vertino PM. GC skew defines distinct RNA polymerase pause sites in CpG island promoters. Genome Res 2015: 25:1600-9.

37. Koch F, Andrau J-C. Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specificity. Transcription 2011; 2:263-8.

38. Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, Farnham PJ. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. Genome Biol 2014; 15:469.

39. Follo M, Finelli C, Mongiorgi S, Clissa C, Chiarini F, Ramazzotti G, Paolini S, Martinelli G, Martelli A, Cocco L. Synergistic induction of PI-PLCβ1 signaling by azacitidine and valproic acid in high-risk myelodysplastic syndromes. Leukemia 2011; 25:271-80.

40. Kalac M, Scotto L, Marchi E, Amengual J, Seshan VE, Bhagat G, Ulahannan N, Leshchenko VV, Temkin AM, Parekh S. HDAC inhibitors and decitabine are highly synergistic and associated with unique gene-expression and epigenetic profiles in models of DLBCL. Blood 2011; 118:5506-16.

41. Soriano AO, Yang H, Faderl S, Estrov Z, Giles F, Ravandi F, Cortes J, Wierda WG, Ouzounian S, Quezada A. Safety and clinical activity of the combination of 5-azacytidine, valproic acid, and all-trans retinoic acid in acute myeloid leukemia and myelodysplastic syndrome. Blood 2007; 110:2302-8.

42. Viré E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, Morey L, Van Eynde A, Bernard D, Vanderwinden J-M. The Polycomb group protein EZH2 directly controls DNA methylation. Nature 2006; 439:871-4.

43. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz MC, Shinkai Y. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 2010; 464:927-31.

44. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res 2012; 22:1128-38.

45. Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, Clark SJ. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. Genome Res 2012; 22:1120-7.

46. Reddington JP, Perricone SM, Nestor CE, Reichmann J, Youngson NA, Suzuki M, Reinhardt D, Dunican DS, Prendergast JG, Mjoseng H. Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. Genome Biol 2013; 14:R25.

47. Liu X, Gao Q, Li P, Zhao Q, Zhang J, Li J, Koseki H, Wong J. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. Nature Commun 2013; 4:1563.

48. Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell 2011; 8:676-87.

49. Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, Jeltsch A. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J Biol Chem 2010; 285:26114-20.

50. Zheng W, Ibáñez G, Wu H, Blum G, Zeng H, Dong A, Li F, Hajian T, Allali-Hassani A, Amaya MF. Sinefungin derivatives as inhibitors and structure probes of protein lysine methyltransferase SETD2. J Amer Chem Soc 2012; 134:18004-14.

51. Edmunds JW, Mahadevan LC, Clayton AL. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. EMBO J 2008; 27:406-20.

52. Carvalho S, Raposo AC, Martins FB, Grosso AR, Sridhara SC, Rino J, Carmo-Fonseca M, de Almeida SF. Histone methyltransferase SETD2 coordinates FACT recruitment with nucleosome dynamics during transcription. Nucl Acids Res 2013:gks1472.

53. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5:R80.

54. Duncan CG, Barwick BG, Jin G, Rago C, Kapoor-Vazirani P, Powell DR, Chi J-T, Bigner DD, Vertino PM, Yan H. A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation. Genome Res 2012; 22:2339-55.

55. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990; 9:811-8.

56. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nature Gen 2006; 38:500-1.

57. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol 2013; 9:e1003118.

Figure Legends

Figure 1: DAC induces global demethylation

A) Average DNA methylation levels after six days (time zero) of DAC treatment. Each circle represents a biological replicate, thick horizontal bars represent the means of all replicates, and error bars are the standard deviation. *, $P < 0.05$, Welch's *t*-test. B) Hierarchical clustering of three untreated and three treated replicate samples across each assayed CpG site. C) Density plot of DNA methylation levels (β values) across all 27K sites in treated and untreated samples at time zero. D,E) A linear mixed effects model (CpG Assoc) was applied to determine CpG sites that were significantly differentially methylated between untreated samples and immediately following treatment with DAC. A change of at least 0.2 in β and an FDR < 0.05 was then imposed to focus on those sites with the greatest biological significance (green). Using these parameters, a total of 5,344 CpGs were significantly demethylated and none were significantly hypermethylated. F) Density plot of DNA methylation levels (β values) across the 5316 CpG sites found to be significantly demethylated in untreated and posttreatment samples at time zero.

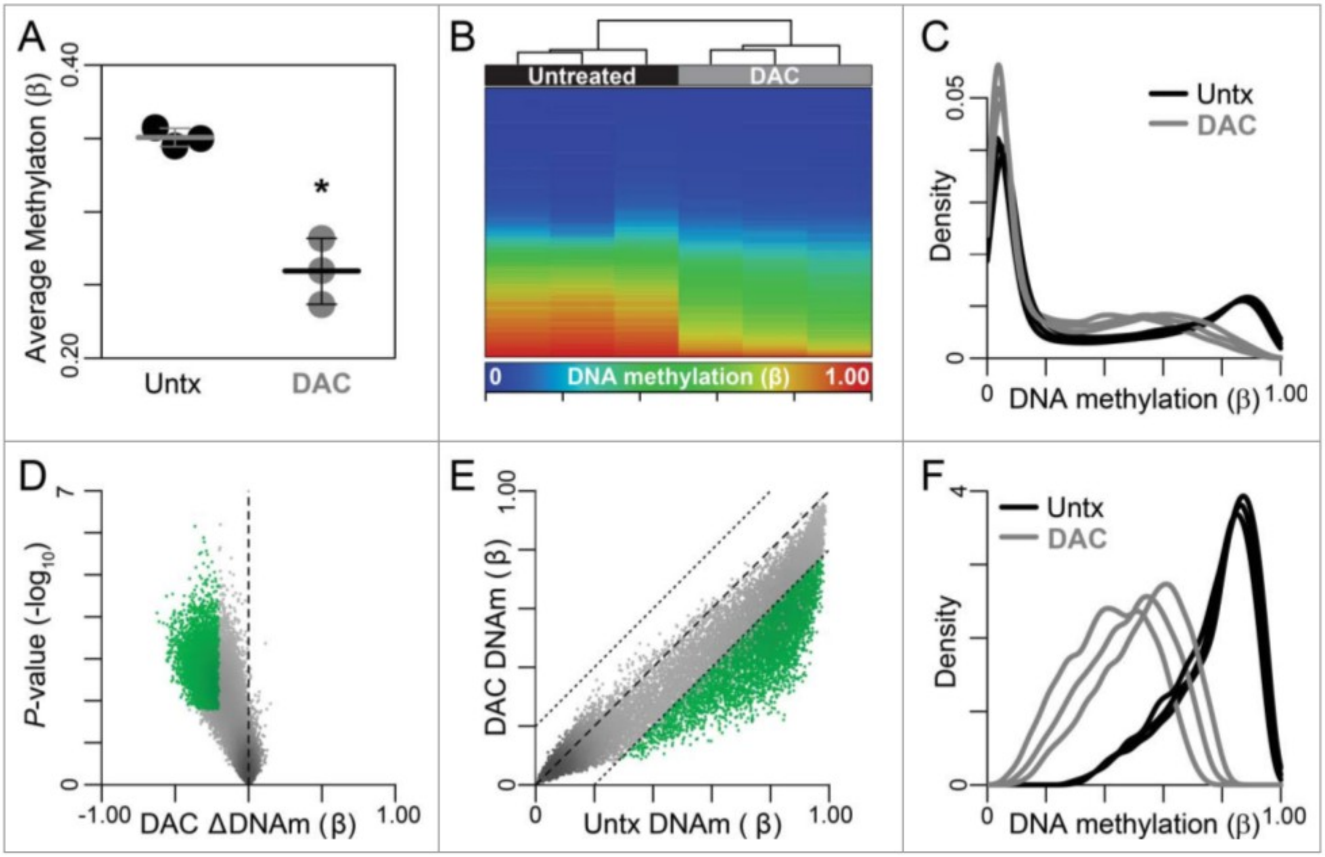

A
0.40
Average Methylaton (β)
0.20
Untx
DAC
*
B
Untreated
DAC
0
DNA methylation (β)
1.00
C
0.05
Untx
DAC
Density
0
DNA methylation (β)
1.00
D
7
P-value (-log$_{10}$)
0
-1.00
DAC ΔDNAm (β)
1.00
E
1.00
DAC DNAm (β)
0
0
Untx DNAm (β)
1.00
F
4
Untx
DAC
Density
0
0
DNA methylation (β)
1.00

Figure 2: Genomic features distinguishing CpG sites significantly demethylated by DAC treatment

A) Distribution of CpGs relative to nearest RefSeq TSS. Demethylated CpGs (N=5316) (red) were plotted by the distance to the nearest TSS oriented to the direction of transcription relative to all probes on the 27K array as a whole (blue). B) Distribution of CGI-associated CpGs (defined as those with a UCSC CpG island (CGI) within 2kb of the nearest RefSeq TSS to the CpG) that underwent DAC-induced demethylation (red) were plotted relative to CpGs on the array as a whole (blue) based on their scaled position within the CGI or their absolute distance to the 5' or 3' edge of the CGI oriented to the direction of transcription. C-H) ChIP-seq data for the indicated histone modification from HMEC cells (ENCODE) was used determine the relationship between chromatin state and demethylation potential. Average normalized tag densities for the indicated histone modification for demethylated CpG sites (red) or all assayed sites on the 27K array (blue) were compiled in 20 bp bins for 5000bp centered on the CpG.

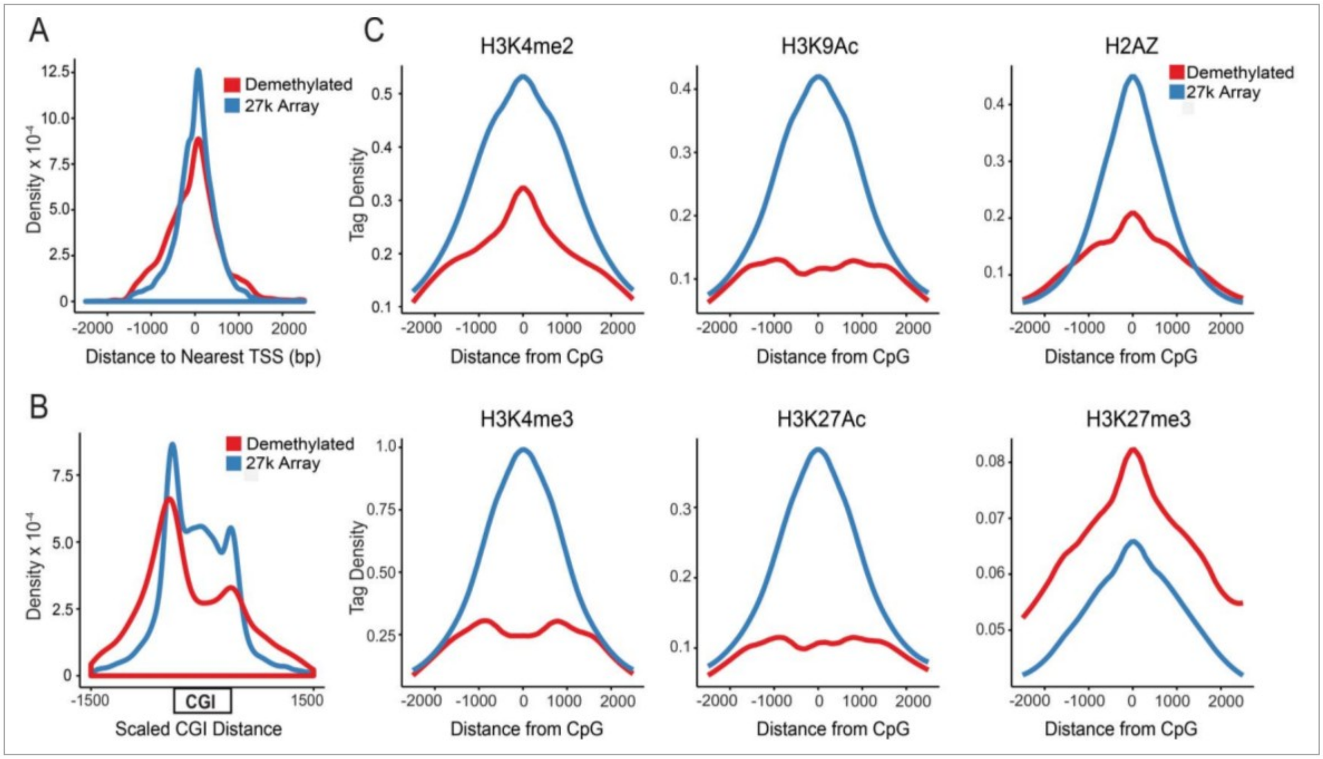

A
B
C
Demethylated
27k Array
Density x 10⁻⁴
Distance to Nearest TSS (bp)
Scaled CGI Distance
CGI
H3K4me2
H3K9Ac
H2AZ
H3K4me3
H3K27Ac
H3K27me3
Tag Density
Distance from CpG

Figure 3: CpG sites demethylated in response to DAC differ in their remethylation kinetics

A) Self-Organizing Maps were used to group significantly demethylated CpGs (n=5316) into 4 classes based on their kinetics of remethylation. Fractional recovery of DNA methylation after DAC-induced demethyation for CpGs in each remethylation class; Resistant, n=1565; Slow, n=1991; Moderate=1244; Rapid=516. Plotted is the median (solid line) and first and third quartiles (shadows) across three biological replicates. B) Relative distribution of DAC-demethylated CpGs in each remethylation class. C) Remethylation kinetics of the same CpGs in an independent experiment in which HCT116 colon cancer cells cells were exposed to 300 μM DAC and allowed to recover in the absence of drug for 42 days (Data from Yang et al,[10]).

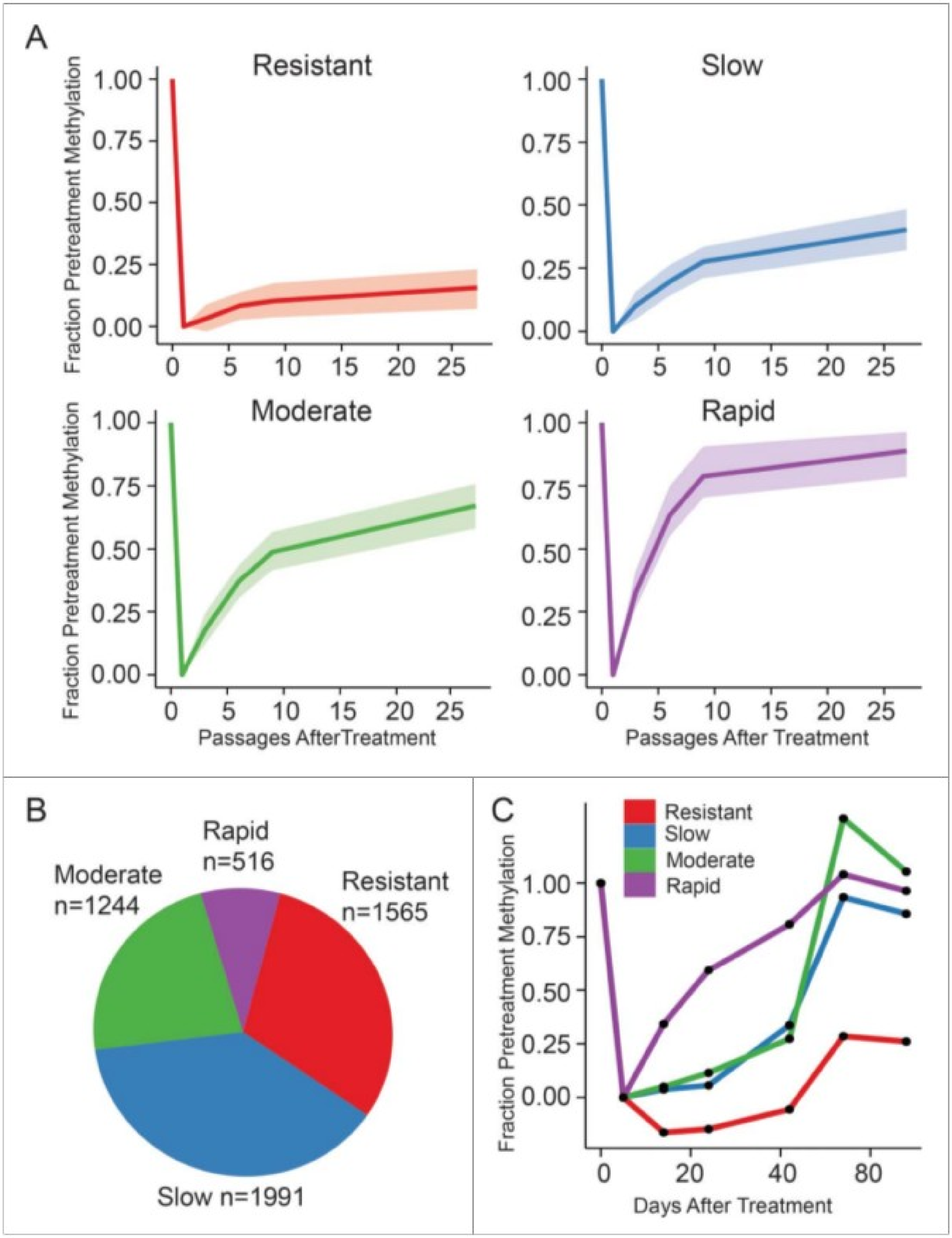

A
Resistant
Slow
Moderate
Rapid
Fraction Pretreatment Methylation
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
Passages AfterTreatment
Passages After Treatment
B
Rapid n=516
Moderate n=1244
Resistant n=1565
Slow n=1991
C
Resistant
Slow
Moderate
Rapid
Fraction Pretreatment Methylation
Days After Treatment
0
20
40
80

Figure 4: Relationship between genomic features with remethylation potential

A) Boxplots representing the distribution of starting methylation levels of CpGs in each remethylation class relative to all demethylated CpGs. The line represents the median, hinges bound the first and third quartiles, and the whiskers represent the maximum or minumum values within 1.5x the interquartile range. B) Fraction of CpGs in and around CGI. Shown is the fraction in each class lying within or relative to a CGI. Shores are defined as 1kb from the edge of the CGI, the North shelf as 1-2.5kb upstream of the CGI, and South Shelf as gene body regions distal to the island (>1kb downstream of the CGI). C) Density plot of CpGs in each remethylation class relative to CGI. Frequency of CpGs were plotted based on the scaled position within a CGI or by absolute distance from either edge of the nearest CGI oriented to the direction of transcription of the nearest TSS. D, E) Density plot of CGI-associated (D) and non-CGI-associated (E) CpGs relative to the position of the nearest TSS. Frequency of CpGs in each class were plotted based on their position to the nearest TSS and oriented to the direction of transcription. CGI- associated sites are defined here as those with a CGI within 2kb of the nearest RefSeq TSS.

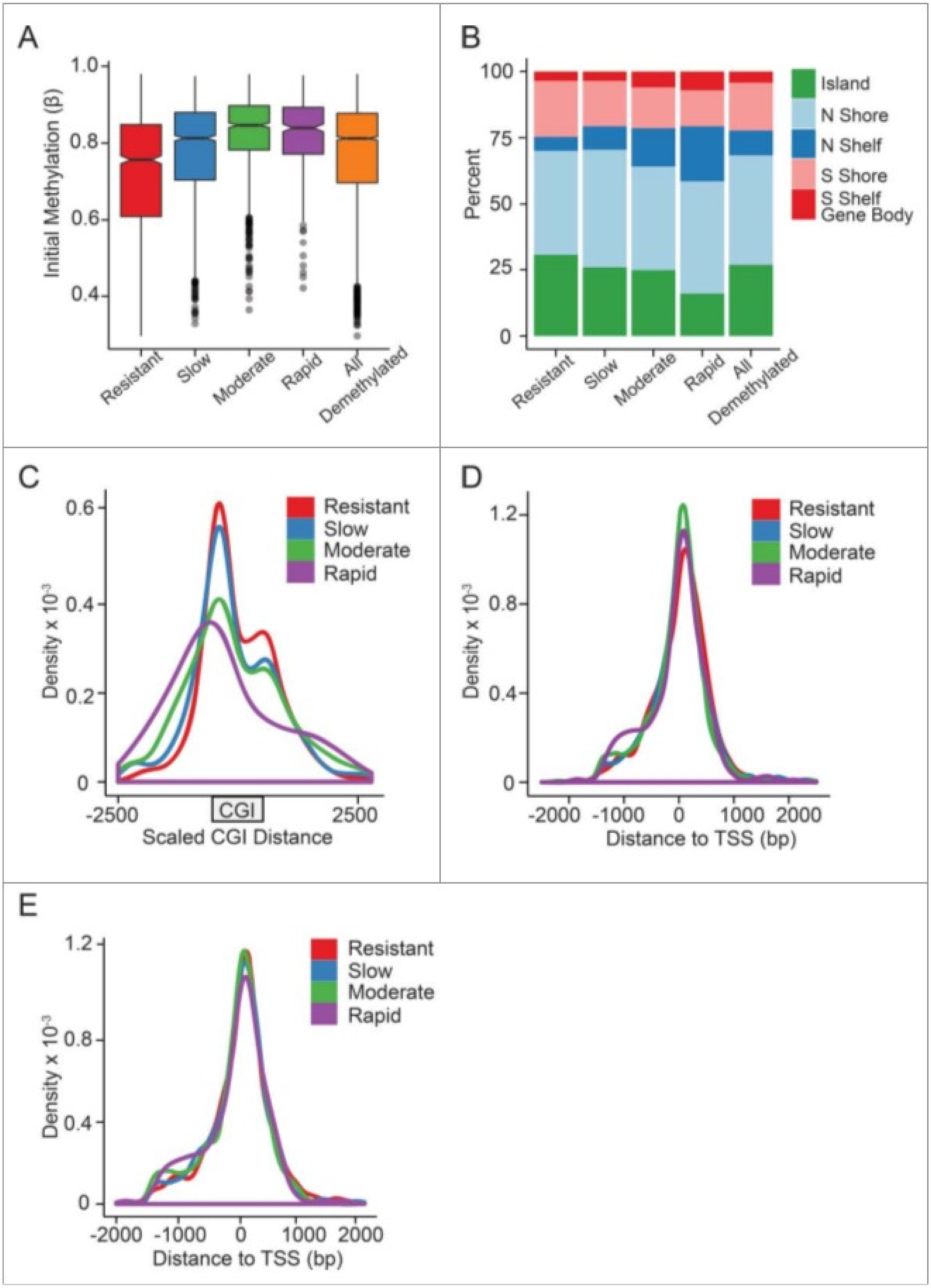

A
Initial Methylation (β)
1.0
0.8
0.6
0.4
Resistant
Slow
Moderate
Rapid
All Demethylated
B
Percent
100
75
50
25
0
Island
N Shore
N Shelf
S Shore
S Shelf
Gene Body
Resistant
Slow
Moderate
Rapid
All Demethylated
C
Density x 10^{-3}
0.6
0.4
0.2
0
Resistant
Slow
Moderate
Rapid
-2500
CGI
2500
Scaled CGI Distance
D
Density x 10^{-3}
1.2
0.8
0.4
0
Resistant
Slow
Moderate
Rapid
-2000
-1000
0
1000
2000
Distance to TSS (bp)
E
Density x 10^{-3}
1.2
0.8
0.4
0
Resistant
Slow
Moderate
Rapid
-2000
-1000
0
1000
2000
Distance to TSS (bp)

Figure 5: Relationship between chromatin functional states and remethylation potential

A) CpGs were assigned to functional chromatin states based on ChromHMM, a Hidden Markov Model based on 9 chromatin modifications. Shown is the fraction of CpG sites in each remethylation class in the indicated ChromHMM category derived from HMEC. B) Odds ratios of enrichment, relative to all demethylated sites, in each chromatin state (*, p<.05, Fisher’s exact). C) Proportion of enhancer associated-CpG sites that represent HMM-defined strong or weak enhancers (Resistant N=231, Slow N=366, Moderate N=209, Rapid N=75).

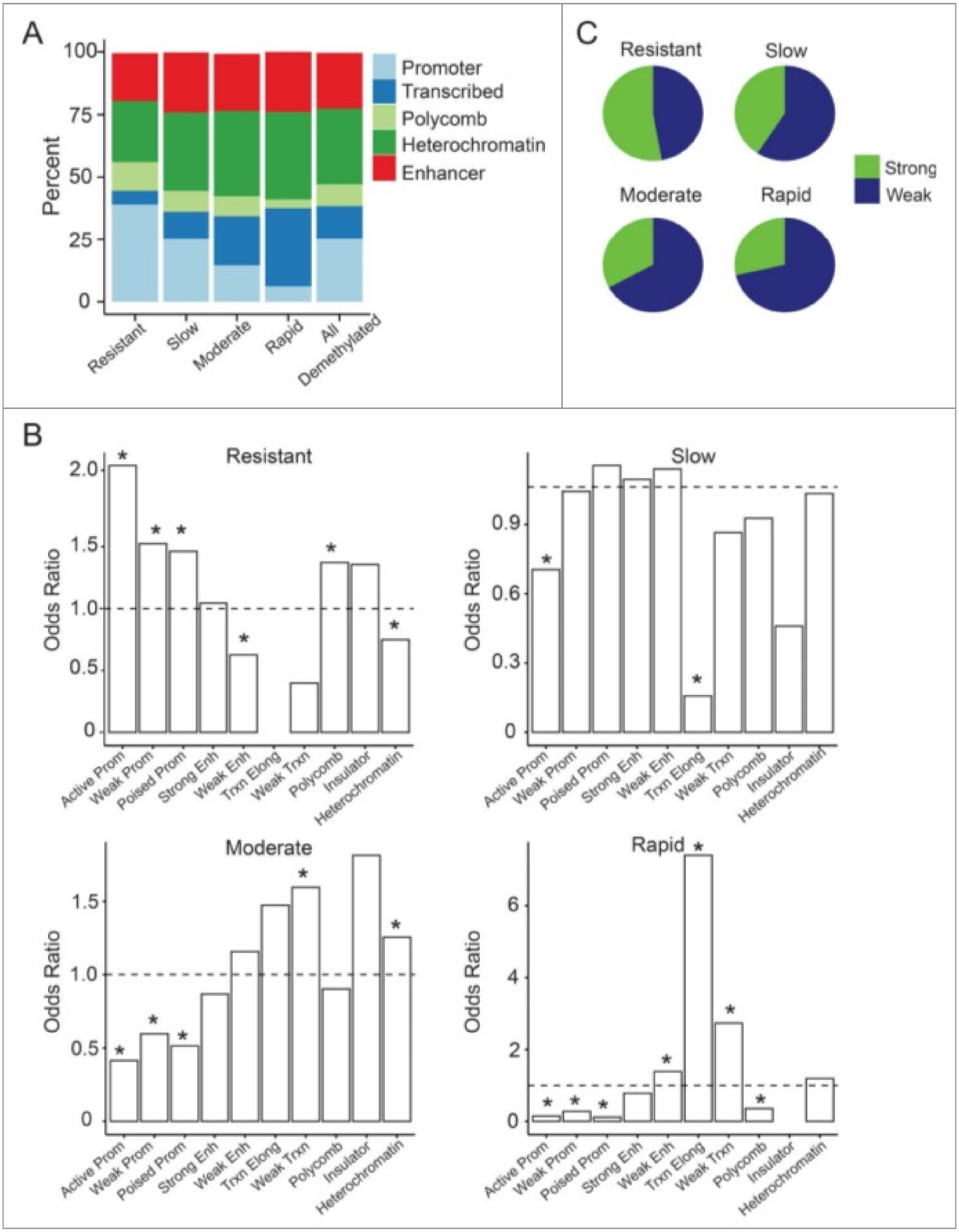

A
100
75
50
25
0
Percent
Promoter
Transcribed
Polycomb
Heterochromatin
Enhancer
Resistant
Slow
Moderate
Rapid
All Demethylated
C
Resistant
Slow
Moderate
Rapid
Strong
Weak
B
Resistant
2.0
1.5
1.0
0.5
0
Odds Ratio
*
Slow
0.9
0.6
0.3
0
Moderate
1.5
1.0
0.5
0
Rapid
6
4
2
0
Active Prom
Weak Prom
Poised Prom
Strong Enh
Weak Enh
Trxn Elong
Weak Trxn
Polycomb
Insulator
Heterochromatin

Figure 6: Histone modification and remethylation potential

ChIP-seq tag densities for the indicated histone modification from HMEC cells (ENCODE) were used to determine the relationship between histone modifications associated with (A) active promoters, (B) gene bodies/ transcribed regions, (C) repressed chromatin, and (D) active enhancers in the regions surrounding CpGs in normal cells and the remethylation potential of the CpG sites. Average normalized tag densities for the indicated histone modification in HMEC cells for CpGs in each remethylation class were compiled in 20 bp bins for 5000bp centered on the CpG.

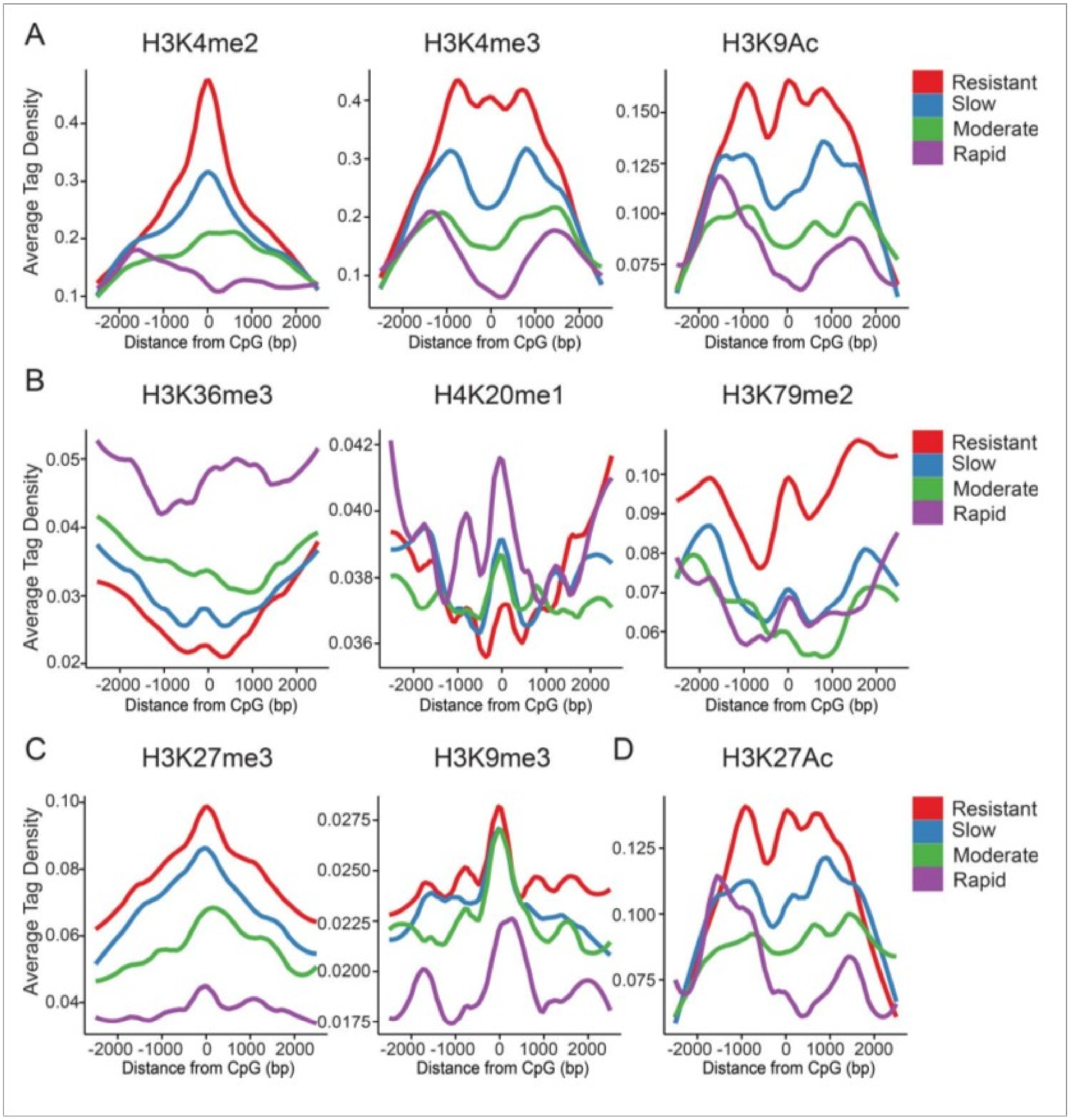
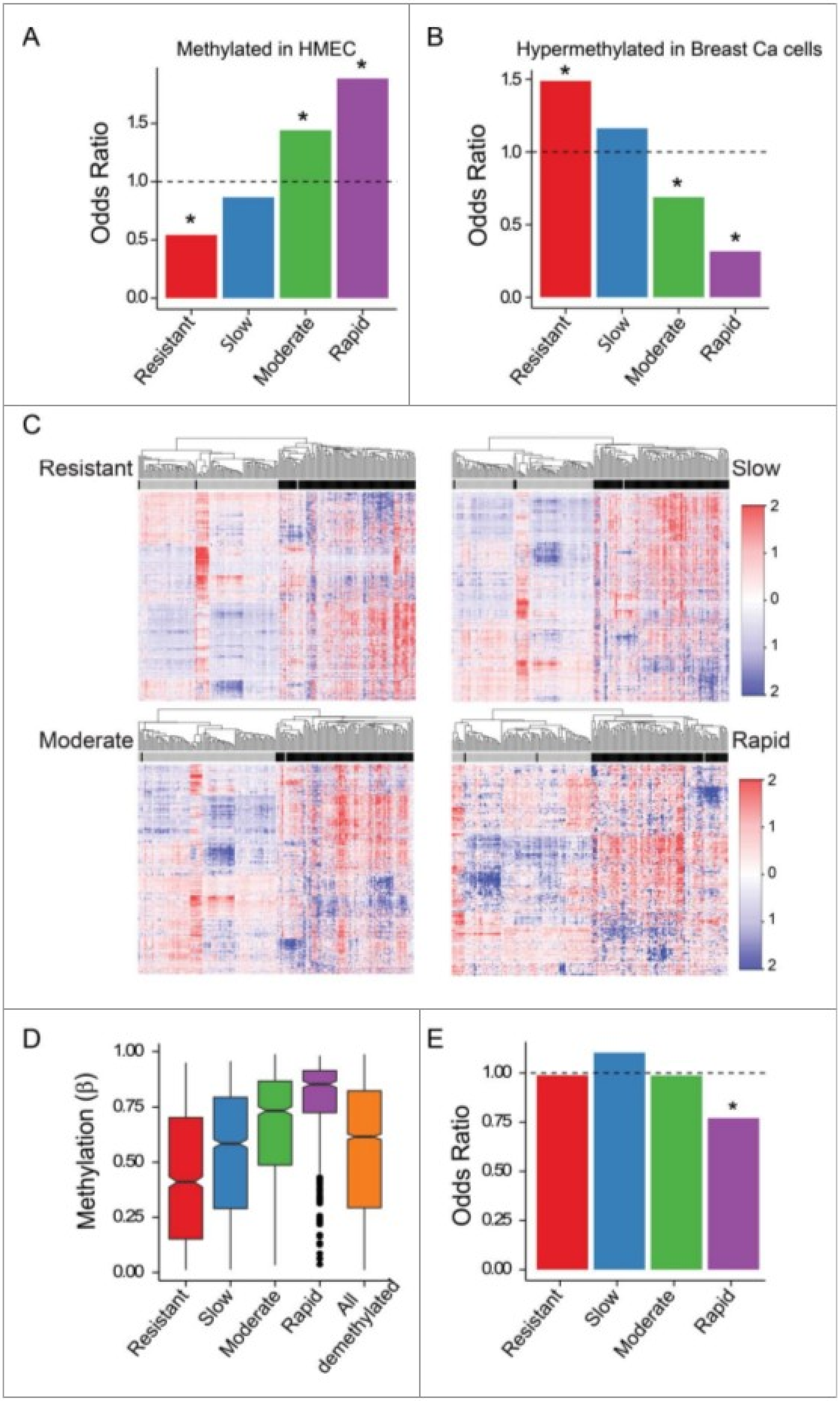

A
H3K4me2
H3K4me3
H3K9Ac
B
H3K36me3
H4K20me1
H3K79me2
C
H3K27me3
H3K9me3
D
H3K27Ac
Average Tag Density
Distance from CpG (bp)
-2000 -1000 0 1000 2000
🟥 Resistant
🟦 Slow
🟩 Moderate
🟪 Rapid
0.1 0.2 0.3 0.4
0.075 0.100 0.125 0.150
0.02 0.03 0.04 0.05
0.036 0.038 0.040 0.042
0.06 0.07 0.08 0.09 0.10
0.04 0.06 0.08 0.10
0.0175 0.0200 0.0225 0.0250 0.0275
0.075 0.100 0.125

Figure 7: CpG sites that undergo cancer-specific hypermethylation versus those normally methylated exhibit distinct remethylation kinetics

A) Enrichment of CpG sites that are endogenously methylated (β > 0.7 in HMEC) in the four remethylation classes. Plotted is the odds ratio (Fisher's exact test). P-values; Resistant= 8.58E-25, Slow=0.0057, Moderate=6.48E-11, Rapid=4.91E-18. B) Enrichment of CpG sites hypermethylated in MB231 cells relative to HMEC (change in β >0.2). Plotted is the odds ratio (Fisher's exact) p-values: Resistant= 1.1E-14, Slow=0.003, Moderate=1.13E-9, Rapid=4.69E-26. C) Methylation status of CpGs in each remethylation class reliably segregate tumor from normal tissues in human breast cancers. The β values for CpGs in each remethylation class were extracted for in 90 matched primary tumor/normal paired breast tissues (TCGA) for the subset of CpGs in each class represented on the Illumina 450k array (Resistant N=1213, Slow N=1527, Moderate N=906, Rapid N=342, total=3988) and used in a hierarchical clustering analysis (agglomerative, complete linkage). D) Box plot representation of the distribution of methylation levels for CpGs in each remethylation class in normal breast tissues (n=90, TCGA). The β values for CpGs in each remethylation class were extracted for the subset of CpGs in each class represented on the Illumina 450k array (Resistant, n=1213; Slow, n=1527; Moderate, n=906; Rapid, n=342). The line represents the median, hinges bound the first and third quartiles, and the whiskers represent the maximum or minimum values within 1.5x the interquartile range. E) CpGAssoc was used to define CpGs significantly hypermethylated in breast cancer (FDR<.05, see methods) and whose methylation status is also represented on the 27K array. This identified 440, 595, 328 and 98 CpG sites in the Resistant, Slow, Moderate, and Rapid classes, respectively. Shown are the odds ratios of enrichment for these hypermethylated sites in each class. Note the underrepresentation of cancer-specific hypermethyation among CpGs in the Rapid class (*, Fisher’s exact, p=0.028).


A
Methylated in HMEC
Odds Ratio
0.0
0.5
1.0
1.5
*
*
*
*
Resistant
Slow
Moderate
Rapid
B
Hypermethylated in Breast Ca cells
Odds Ratio
0.0
0.5
1.0
1.5
*
*
*
Resistant
Slow
Moderate
Rapid
C
Resistant
Slow
Moderate
Rapid
2
1
0
1
2
2
1
0
1
2
D
Methylation (β)
0.00
0.25
0.50
0.75
1.00
Resistant
Slow
Moderate
Rapid
All demethylated
E
Odds Ratio
0.00
0.25
0.50
0.75
1.00
*
Resistant
Slow
Moderate
Rapid

Supplemental Data and Methods

Supplemental Methods:

COmbined Bisulfite Restriction Analysis (COBRA)

COBRA analysis was performed as previously described (Kagey et al. 2010). Briefly, bisulfite-modified DNA was amplified using primers devoid of any CpGs. Amplified products were purified with the PCR Purification kit (Qiagen), digested overnight at 37 °C with either *Bss*HII, *Mlu*I, *Xmn*I or *FNU*4HI, precipitated, and resolved on a 2.0% agarose gel. Relative intensities of digested and undigested bands were quantified with Image Quant 5.2 and percent methylation was determined as the combined intensity of the digested bands relative to that of all bands (undigested and digested). Primer sequences, restriction enzyme, and the genomic location of the queried CpG for each locus is listed in Supplemental Table I.

GSEA/ GO Analysis

Gene sets from each remethyaltion class were analyzed for overlap with curated data sets (C2, C4, C6, C7, H, MF) in MSigDB using the web interface available at http://www.broadinstitute.org/gsea/msigdb/ (Subramanian et al. 2005) and for functional annotation using the DAVID Bioinformatics Resource (http://david.abcc.ncifcrf.gov) (Dennis et al. 2003). Input gene lists (symbols) were created by annotating each CpG site to the nearest RefSeq gene within 5kb. Only genes uniquely assigned to a single remethylation class were considered. These data can be found in Supplemental File GSEA_GO.xls, and labelled therein as individual worksheets (eg. GSEA_RAPID, GO_ Rapid, and so on).

Supplemental References:

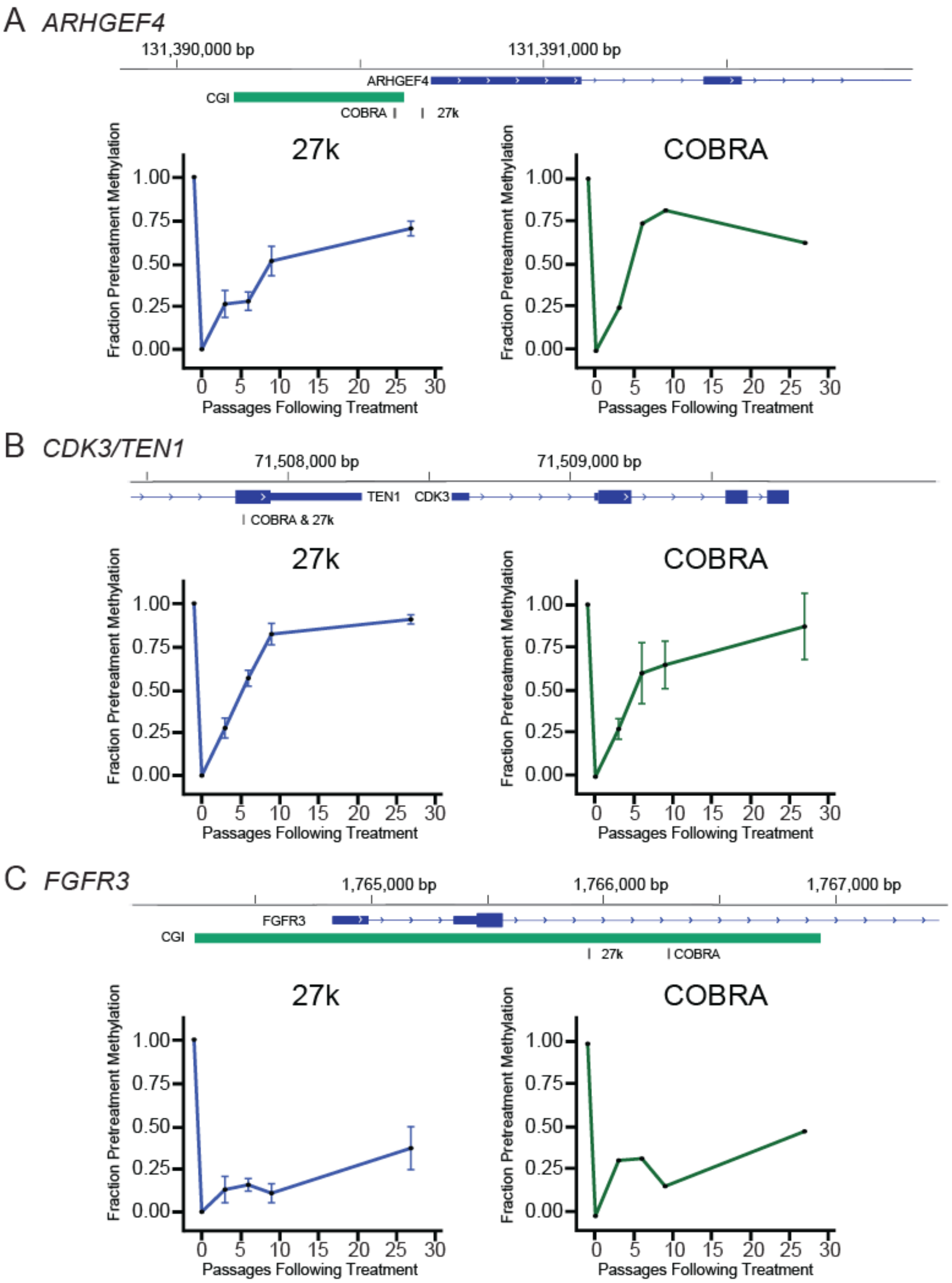
Dennis G, Sherman, BT, Hosack, DA, Yang J, Baseler BW, Lane HC, Lempicki, RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. 2003; *Genome Biol* 4: P3

Kagey JD, Kapoor-Vazirani P, McCabe MT, Powell DR, Vertino PM. Long-term stability of demethylation after transient exposure to 5-aza-2′-deoxycytidine correlates with sustained RNA polymerase II occupancy. *Mol Cancer Res* 2010; **8**:1048-59.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A.* 2005; 102:15545-50.

.

Supplemental Figure and Legend:



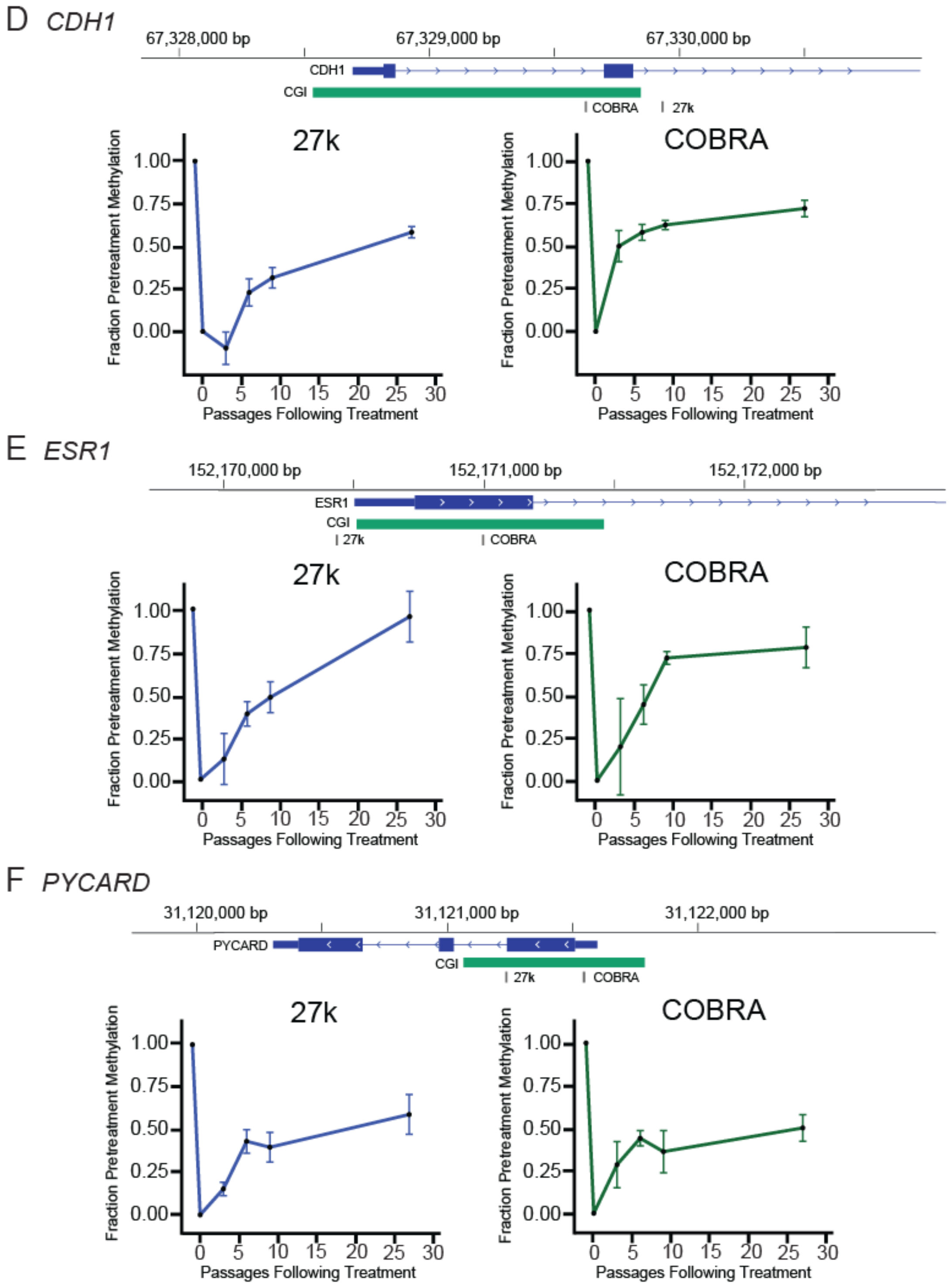

D CDH1
67,328,000 bp
67,329,000 bp
67,330,000 bp
CDH1
CGI
COBRA
27k
27k
COBRA
Fraction Pretreatment Methylation
1.00
0.75
0.50
0.25
0.00
0 5 10 15 20 25 30
Passages Following Treatment
E ESR1
152,170,000 bp
152,171,000 bp
152,172,000 bp
ESR1
CGI
27k
COBRA
27k
COBRA
Fraction Pretreatment Methylation
1.00
0.75
0.50
0.25
0.00
0 5 10 15 20 25 30
Passages Following Treatment
F PYCARD
31,120,000 bp
31,121,000 bp
31,122,000 bp
PYCARD
CGI
27k
COBRA
27k
COBRA
Fraction Pretreatment Methylation
1.00
0.75
0.50
0.25
0.00
0 5 10 15 20 25 30
Passages Following Treatment

Figure S1: Comparison of remethylation kinetics of CpG sites in the *ARHGEF4*, *CDK3/TEN1*, *FGFR3* loci (A-C) or the *CDH1*, *ESR1*, *PYCARD* loci (D-F) as determined by the Illumina 27k array or by COBRA analysis

The fractional recovery of DNA methylation after 6 days DAC-induced demethylation was determined for CpGs as measured on the Illumina 27k array (left) or by COBRA (right). Shown are IGV browser images depicting the relative position of the CpG sites interrogated on the Illumina array (27K) or by COBRA analysis (COBRA) relative to the nearest RefSeq gene(s) and CpG island (see Supplemental Table I). For *CDK3*, the same CpG site is assayed by both methods. (A-F) The 27K array data represent the mean +/- sd of Beta values as determined from three independent time course experiments as described in the Methods section. DNA from the same time courses were separately analyzed by COBRA. (A) COBRA data represent the mean +/- sd of single determinations from the three independent time courses. (B,C) For *ARHGEF4* and *FGFR3*, the COBRA analysis was determined from a single time course experiment. (D-F) For comparative purposes, raw COBRA data for *CDH1*, *ESR1*, and *PYCARD* were extracted from Kagey et.al. 2010 and re-analyzed to determine the percent recovery after drug removal. Data represent the mean +/- s.d. of single determinations from the same three independent time courses reported in this study.

Supplemental Table I – Genomic location of CpG sites interrogated by Illumina 27K platform and COBRA Analysis (see Figure S1)

Gene	Remethylation Class	Genomic Location* (27K CpG)	COBRA Primers	COBRA Enzyme	Genomic Location* (COBRA CpG)
CDK3	Rapid	chr17:71507841	TATTGGGAGTTTAGTTTTTTGG CTACTATTTCCTACTAACTACC	BssHII	chr17:71507841
FGFR3	Slow	chr4:1765940	GATTTTTAAGGGTGGGTGTG AACCAAAACCTCCCTCCAC	MluI	chr4:1766279
ARHGEF4	Moderate	chr2:131390671	AGAGTTTGGGAGAGTGTTGG CCAAAATCCCCTAAAATCCCC	Xmn1	chr2:131390591
PYCARD	Moderate	chr16:31121231	TTGGTGTAAGTTTAGAGATAAGT ACCATCTCCTACAAACCCATA	Fnu4HI	chr16:31121541
CDH1	Slow	chr16:67329931	GAGGGAAGGAGAGGGGTATT CCCTCACCTCTACCCAAAAC	Fnu4HI	chr16:67329626
ESR1	Moderate	Chr6:152170436	AGGTGTATTTGGATAGTAGTAAG CAAATAATAAAACACCTACTAACC	Fnu4H1	chr6:152171000

*Genomic coordinates are Hg18

Chapter IV: Orphan CpG Islands Define a Novel Class of Highly Active Enhancers

Joshua S.K. Bell & Paula M. Vertino

This work is published in *Epigenetics* 2017 Jun 3;12(6).

Contributions: JB designed and performed the analysis and wrote the manuscript.

Abstract

CpG islands (CGI) are critical genomic regulatory elements that support transcriptional initiation and are associated with the promoters of most human genes. CGI are distinguished from the bulk genome by their high CpG density, lack of DNA methylation, and euchromatic features. While CGI are canonically known as strong promoters, thousands of 'orphan' CGI lie far from any known transcript, leaving their function an open question. We undertook a comprehensive analysis of the epigenetic state of orphan CGI across over 100 cell types. Here we show that most orphan CGI display the chromatin features of active enhancers (H3K4me1, H3K27Ac) in at least one cell type. Relative to classical enhancers, these enhancer CGI (ECGI) are stronger, as gauged by chromatin state and in functional assays, are more broadly expressed, and are more highly conserved. Likewise, ECGI engage in more genomic contacts and are enriched for transcription factor binding relative to classical enhancers. In human cancers, these epigenetic differences between ECGI versus classical enhancers manifest in distinct alterations in DNA methylation. Thus, ECGI define a class of highly active enhancers, strengthened by the broad transcriptional activity, CpG density, hypomethylation, and chromatin features they share with promoter CGI. In addition to indicating a role for thousands of orphan CGI, these findings suggests that enhancer activity may be an intrinsic function of CGI in general and provides new insights into the evolution of enhancers and their epigenetic regulation during development and tumorigenesis.

Introduction

Vertebrate genomes are heavily methylated, CpG-poor, predominantly heterochromatic terrains disrupted by CpG Islands (CGI), essential CpG dense regulatory elements. CGI are defined by a lack of DNA methylation, transcriptional competence, and heightened euchromatic features [1], such as enrichment of H3K4me3 and H3K9/27Ac. These regions are found at the promoters of nearly two-thirds of human protein-coding genes. As such, CGI have been canonically studied for their role in permitting transcriptional initiation. CGI-associated promoters demonstrate broader expression patterns across tissues and tend to be stronger than CpG-poor promoters [2]. These properties are contingent on remaining unmethylated, as hypermethylation of CGI leads to the recruitment of repressive complexes containing histone deacetylases and chromatin remodelers [3] resulting in the silencing of the associated gene [4-6]. This is a cardinal method by which cancer cells inactivate tumor-suppressor genes [7-9].

CpG sites are targeted for DNA methylation during development [10] unless actively protected by methylation of histone H3 lysine 4 (H3K4), a mark found principally at promoters and enhancers and associated with transcriptional initiation [11]. Methylated cytosine is mutagenic due to a propensity to undergo deamination to thymine, leading to the paucity of CpG sites throughout the genome [12] and marking CpG-rich regions as being of potential functional importance. Yet, half of CGI do not in fact overlap known TSSs and thousands lie far from any known transcript. While many of these 'orphan' CGI are likely promoters of unannotated transcripts, their prevalence suggests there may be roles for CGI other than as strict promoters.

Enhancers are regulatory elements that act at a distance to promote gene transcription independent of position or orientation. Enhancers are similar to CGI in that their levels of DNA methylation are inversely correlated with their activity [13; 14]. We also recently demonstrated that following treatment with decitabine, a chemotherapeutic demethylating agent, strong enhancers and promoter CGI are more resilient to *de novo* methylation than weak enhancers[15]. Indeed, recent work has demonstrated that enhancers and promoters possess a unified chromatin architecture, characterized by the presence of H3K4me, H3K9/27Ac, DNAse hypersensitivity, evidence of paired bidirectional transcriptional initiation, and transcription factor binding. Measures of transcript stability provide the most reliable method to distinguish the two, with promoters giving rise to one or two stable transcripts, and enhancers strictly unstable (eRNA) transcripts[16].

The striking similarity of enhancers and promoters suggests that CGI, especially those without evidence of a nearby genes may act as enhancers [17; 18]. Here, we undertake a comprehensive analysis of orphan CGI chromatin topology across over one hundred cell types. We demonstrate that the vast majority of orphan CGI appear to be active enhancers in at least one cell type. These enhancer CGI (ECGI) are much more powerful than classical enhancers in their ability to drive transcription by a variety of measures, manifesting open chromatin across a broader variety of cell types, with heightened genomic contacts, and stronger enrichment for transcription factor binding. These features contribute to the evolutionary conservation of CpG density relative to other enhancers and to distinct susceptibilities to alterations in DNA methylation in cancer.

Results

Orphan CpG Islands Exhibit Features of Active Enhancers

Given the established role for CGI in promoting transcriptional initiation, we first sought to characterize the relationship between CGI across the genome with known transcripts, focusing on UCSC CGI [19] (N= 27,718) and the GENCODE (V25) transcript database [20; 21](Fig 1A). Strikingly, only 45% (N=12,548) of CGI contain an annotated TSS for a protein-coding gene. In contrast, 32% (N=8899) are found within a protein-coding gene, and 3% (N=920) are within 2kb of, but do not directly overlap a gene (perigenic). An additional 6% of CGI overlap or are within 2kb of noncoding (ncRNA) transcripts (N=1131 long noncoding (ncRNA), N=837 other ncRNA, see methods), and 2.5% are found near or overlapping pseudogenes. The 10% of remaining CGI (N=2693), we term 'orphan' CGI because they cannot be annotated to any known transcript (are more than 2kb from any known transcript). This distribution suggests that many or most CGI in the genome may not be acting strictly as promoters.

Given that enhancers and promoters are characterized by a similar epigenetic environment we hypothesized that orphan CGI may be enhancers. In order to test this, we defined putative enhancers using two established methods: regions containing overlapping peaks of H3K4me1 and H3K27Ac [22; 23] across 120 cell types (22 ENCODE [24], 98 Roadmap Epigenomics Project [25], see Methods), as well as regions annotated as enhancers in chromatin state maps based on Hidden Markov Models (chromHMM [26]) in 136 cell types (9 ENCODE, 127 Roadmap). Using these definitions, fully 92% of Orphan CGI overlapped an enhancer in one or more cell types (2241 overlapped enhancers by both definitions, 197 peak only, 33 HMM only, Fig.1B,C).

Given the prevalence with which orphan CGI can confidently be called enhancers, we investigated whether other classes of CGI (containing a TSS or otherwise transcript-associated) could be classified as enhancers (Supplementary Fig. 1A). We find that promoter CGI are especially likely to contain H3K4me1/H3K27Ac peaks (in ~60% of cell lines examined, Supplemental Fig. 1A), although each class, especially perigenic CGI, were likely to contain enhancer peaks in at least a quarter of cell lines. We found no evidence of enhancer activity for 8% (N=227) of Orphan CGI, which we term 'Remnant' CGI. Overall, Orphan CGI overlapped an HMM-defined enhancer in a median of 15 cell types or by H3K4me1/H3K27Ac peak overlap in 20 cell types. We term those Orphan CGI that overlap regions that satisfy both enhancer criteria in at least one cell type (n=2241) Enhancer CGI or ECGI.

To study ECGI in greater detail, we focused on those that overlapped both an H3K4me1/H3K27Ac peak and an HMM-defined enhancer in one of three cell lines: H1 human embryonic stem cells (H1ESC, N=180), human mammary epithelial cells (HMEC, N=205), and K562 (N=169), a human myelogenous leukemia line. We chose these lines for their phenotypic diversity (stem cell, normal differentiated cell, and cancer cell), and because each is well-studied with an abundance of publicly-available epigenomic datasets. For comparison, we defined non-CGI classical enhancers in each cell line by the same criteria (H3K4me1/H3K27Ac peaks that overlap HMM enhancers, N=11863 HMEC, N=12138 H1ESC, N=7610 K562), as well as CGI overlapping the TSS of protein-coding genes, apparent canonical promoter CGI (N=12548).

Distinguishing promoters from enhancers is not straightforward, however Core et al. [16] recently characterized initiation regions in mammalian genomes by comparing the TSSs of stable transcripts detected by CAGE, which captures 5’ 7-methylguanylate capped, steady-state transcripts like mRNAs or lncRNAs, versus those detected by GroCap, which detects the TSSs of all nascent transcripts, including unstable ones like eRNAs or upstream antisense RNAs (uaRNA). They found that enhancers could be defined by unstable transcript pairs, and promoters by a stable transcript paired with either an unstable uaRNA or another stable transcript. Focusing on ECGI and non-CGI classical enhancers active in K562 cells (Fig. 1D), we found that ECGI, like classical enhancers, exhibit strong enrichment for unstable-unstable pairs, in stark contrast to promoter CGI and other non-CGI promoters (Gencode TSSs >2.5kb from a CGI) which tend to be enriched in stable-stable or stable-unstable pairs (for unstable-unstable pairs in ECGI vs. promoter CGI: Odds Ratio (OR)=6.36, p=6.4E-19, Fisher's Exact). Indeed, no other class of transcript-associated CGI exhibited the preponderance of putative eRNA (unstable) pairs displayed by orphan CGI (Supplemental Fig. 1B,C). Intragenic, perigenic, and ncRNA CGI were equally likely to contain unstable as stable transcripts, which likely mark alternative promoters, a documented role of CGI distinct from enhancer function [27], whereas promoter CGI were enriched in stable pairs as expected (Supplemental Fig. 1B,C). Notably, CGI associated with known pseudogenes were unlikely to exhibit detected transcript pairs at all, or to overlap enhancer chromatin peaks, consistent with their transcriptional inactivity. Together, the prevalence of enhancer chromatin features and unstable transcripts suggest that many CGI lying near or within known extant transcripts may also exhibit enhancer activity. These data are also robust evidence that ECGI are not simply unannotated promoters, but are in fact enhancers.

DNA hypomethylation is a key feature of promoter CGI, but is also linked to the activity of enhancers [13; 14]. To assess DNA methylation patterns at ECGI, we utilized a Whole Genome Bisulfite Sequencing (WGBS) data from normal breast tissue (TCGA [28]) to compare the average levels of DNA methylation at promoter CGI, ECGI, and enhancers as defined in HMECs (Fig 1N). We find that like promoter CGI, active ECGI display minimal DNA methylation (<10%), while classical enhancers exhibit much more variable methylation, typically 50-80%. Because DNA hypomethylation is intrinsically linked to CpG density, we quantified the GC content and CpG density of ECGI (Fig. 1F). While ECGI and promoter CGI have similar GC content (median 70% G+C for both, p=0.45), ECGI are slightly less CpG dense (median 0.82 observed/expected for ECGI vs median 0.86 for promoter CGI p=4.28E-

9). However, both CGI classes exhibit far higher GC content and observed/expected CpG density than do classical enhancers.

Given the ways DNA methylation is known to affect histone modifications and other epigenetic features, we next examined the chromatin state of ECGI relative to promoter CGI and other enhancers. Focusing initially on the endogenous chromatin state of features defined in HMEC cells, we found that ECGI tend to have overall higher levels of H3K27Ac than either classical enhancers or promoter CGI (Fig. 1G,K), suggestive of highly active chromatin. Levels of H3K4 methylation have been used to distinguish enhancers from promoters, with enhancers defined as having high levels of H3K4me1, and promoters H3K4me3. However, neither is exclusive as the strongest enhancers exhibit H3K4me3 and it is the ratio of H3K4me3 to H3K4me1 that has been tightly correlated with transcriptional intensity [16]. Consistent with this idea, ECGI, like classical enhancers, exhibit substantial H3K4me1 (Fig. 1H,L), which is absent in promoter CGI. Interestingly, ECGI uniquely possess abundant H3K4me2 (Fig. 1 IM), a less-studied mark usually linked to the transition between H3K4me1/3 [29; 30]. ECGI also display modest levels of H3K4me3 (Fig. 1J,N) and an intermediate ratio of H3K4me3/H3K4me1 (Fig. 1O): much greater than that of classical enhancers, but lower than that of canonical promoter CGI. Together, these data suggest that ECGI display a chromatin state similar to that of the most active enhancers, and distinct from that of promoter CGI.

ECGI are Stronger than Classical Enhancers

The finding that ECGI exhibit a higher H3K4me3/H3K4me1 ratio and less DNA methylation than classical enhancers suggests that they may be more active than classical enhancers. To investigate this relationship, we ascertained the levels of other features often used to gauge enhancer activity: GroSeq tag density, a measure of nascent transcription or eRNA production [31; 32], enrichment of H3K27Ac, H3K9Ac, and DNase hypersensitivity, a measure of open chromatin (Fig. 2A-D). We find that ECGI display much stronger enrichment for each of these features of activity than do classical enhancers [ECGI vs. classical enhancers: GroSeq, p= 6.1E-22; H3K27Ac, p=3.48E-16; H3K9Ac, p=2.48E-54; DNase hypersensitivity, p=1.72E-7; Mann-Whitney U].

Super enhancers represent a powerful subset of enhancers that exhibit the greatest genomic enrichment of features critical to enhancer function: typically defined by H3K27Ac levels or the degree of binding of transcriptional co-regulators like Mediator and BRD4, among others [33]. Utilizing the Super Enhancer Archive (SEA) database [34], we found that ~20% of ECGI are putative super enhancers, compared to less than 10% of classical enhancers and less than 5% of promoter CGI (combined ECGI vs. Enhancers, OR= 3.84, p= 1.02E-30, Fishers exact) (Fig. 2E).

The above data suggest that ECGI represent a subset of enhancers distinct from classical enhancers in terms of strength. Thus, we next sought to ascertain the functional enhancer activity of ECGI. Inoue et al. [35] recently screened over 2000 putative enhancer elements by cloning them into a GFP enhancer-reporter vector capable of integrating into the genome in an assay known as lentiviral Massively Parallel Reporter Assay, in which the strength of such elements is determined by comparing the GFP reporter mRNA levels with the DNA copy number in transfected cells. We compared the activity of ECGI, other CGI, and other putative enhancers tested in the assay, and found that both ECGI (N=13) and CGI in general (N=171), exhibited much stronger ability to enhance transcription than other elements (N=2055) [Fig. 2F; ECGI vs. non-CGI elements, p=0.00051, other CGI vs. non-CGI elements, p=3.26E-20, ECGI vs. other CGI p=0.2]. We also examined two additional functional enhancer screens conducted in mouse cells, FIREWACh [36] and CapStarr-seq [37] (See Supplemental Data). In both assays, we found that mouse CGI in general, but especially those conserved as human ECGI, exhibit stronger enhancer activity than classical enhancers (Supplemental Fig. 2). However, human ECGI that have lost their CpG density in the mouse showed a much reduced ability to enhance transcription. These results indicate that while CGI in general can exhibit potent enhancer activity, ECGI in particular are functionally stronger than classical enhancers, dependent upon the conservation of their CpG density.

ECGI are more broadly active than typical enhancers

We next addressed the degree of cell-type specificity exhibited by ECGI vs. classical enhancers by comparing the fraction of tested cell lines in which each feature also exhibited marks of active enhancers (H3K4me1/H3K27Ac peaks). We found that the ECGI in each cell line (H1ESC, HMEC, K562) was active in a median of 50-75% of cell lines examined, compared to a median of just 25-30% of typical enhancers (Fig. 3A, combined ECGI vs. enhancers p= 6.39E-116; Mann-Whitney U).

Moreover, ECGI and classical enhancers defined in HMEC exhibit the highest levels of each feature of enhancer activity: H3K4me1, H3K27Ac, H3K9Ac, and DNAse hypersensitivity in HMEC cells, as expected (Fig. 3B-E). Yet, H1ESC and K562-derived ECGI also exhibit significant enrichment for each of these features in HMEC cells, consistent with constitutive activity for a substantial proportion. In contrast, classical enhancers from these lines tend to lose these features and appear inactive in HMEC cells, consistent with more cell-type restricted activity. Consistent with this idea, both H1ESC and K562-defined ECGI as well as classical enhancers were enriched in H3K27me3 in HMEC relative to the cell type of origin (Fig. 3F), demonstrating Polycomb-mediated repression. However, in contrast to classical enhancers, ECGI were unlikely to be marked by H3K9me3 (Fig. 3G), consistent with their DNA hypomethylation. We found similar trends towards cell type restricted activity when comparing the chromatin state of ECGI vs. classical enhancers in H1ESC and K562 cells (Supplementary Figure 3). These data indicate that ECGI are more likely than classical enhancers to be active across multiple cell types. When they do undergo silencing, ECGI and classical enhancers both exhibit H3K27me3, but only classical enhancers appear prone to H3K9me3-mediated repression.

ECGI are hubs of genomic contacts

The ability of enhancers to form physical loops with gene promoters is a critical feature of their activity [38]. We find that ECGI exhibit greater enrichment than classical enhancers for proteins involved in enhancer/promoter contacts including CTCF [39], Cohesin [40], and BRD4 [41]; [42] (Fig. 4A) [Combined p-value for ECGI in all three lines vs. classical enhancers in all three lines CTCF p=1.44E-12, Cohesin p=4.45E-18, BRD4 p=7.11E-90]. Using K562 Hi-C data, which detects all physical contacts in an unbiased manner [43], we find that ECGI, like promoter CGI, exhibit a much greater contact frequency than do classical enhancers (Fig. 4B), regardless of the cell line in which they are active. However, ECGI defined in K562 exhibited the strongest enrichment, indicating modest cell-type specificity (combined ECGI vs. enhancers p= 5.52E-129, K562 ECGI vs. other ECGI p=2.21E-5, Mann-Whitney U). Similar results were observed in CTCF and RNA Polymerase II ChiAPet data (ENCODE), which exposes only contacts between fragments containing these factors and thus more likely to be of functional relevance [44] [combined CTCF OR=5.12, p=1.96E-51, Pol II OR=4.48 & p=1.55E-56, Fisher's exact, Fig. 4C].

Topologically Associating Domains (TADs) are large genomic regions within which physical interactions are likely to occur. TADs are vital to gene regulation and nuclear organization, often serving to partition active from inactive chromatin [45]. Because CTCF and Cohesin are important to the formation of TADs, we examined the distance to TAD boundaries for ECGI and classical enhancers previously determined in high resolution Hi-C studies in GM12878 cells [43]. We find that ECGI, like promoter CGI, on average lay much closer to TAD boundaries than do classical enhancers (combined ECGI vs. enhancers p=2.06E-30, Mann Whitney U, Fig. 4D), suggesting a role for ECGI in global nuclear organization, or a requirement for tight regulation of their chromatin state.

ECGI are enriched in GC and CpG-rich transcription factor binding sites

Enhancers are enriched in transcription factor (TF) binding sites and TF binding is correlated with their activity and chromatin state [33; 46], leading us to investigate the identity and degree of TF binding at ECGI. The JASPAR database contains approximately 1.1 million annotated binding sites for ~130 TFs based on binding motif sequence [47]. We find that ECGI, like promoter CGI, contain ~2-4 potential TF binding sites per kb, while most classical enhancers had one or no sites (Fig. 5A, combined p=2.16E-38, Mann-Whitney U). Overall, binding sites for 33 TFs were significantly enriched (OR>1, p<.05) in ECGI relative to classical enhancers (Fig. 5B). Strikingly, we find that the top 7 most enriched TF motifs (SP2, E2F4, E2F1, NRF1, ZBTB33, E2F6, and EGR1) had GC contents greater than 50% and contained a CpG site, leading us to quantify the association between these intrinsic features of CGI and TF motif density. Consistent with the high GC content and CpG density of CGI in general, especially relative to that of classical enhancers (Fig.1F), we find that there is a direct correlation between the GC content of a motif and its relative enrichment in ECGI vs classical enhancers (Pearson correlation ρ=0.54, p=.00172 , Fig. 5C). Furthermore, even within a GC-rich context, motifs containing a CpG were much more likely to be enriched in ECGI vs. classical enhancers (median CpG-motif OR=62.5 vs. median non-CpG-motif OR 5.2, p=2.5E-4 Mann-Whitney U, Fig 5D). To determine whether this enrichment of GC/CpG-rich motifs corresponded to the degree of actual TF binding to chromatin in cells, we utilized ENCODE ChIP-Seq data performed in K562 cells to compare binding of 6 of the top-scoring TFs at K562-specific ECGI and classical enhancers. Binding for each of these factors was strongly enriched at ECGI relative to classical enhancers [SP1, p=2.18E-26; SP2, p=6.04E-35; EGR1, p=3.89E-71; NFYA, p=1.06E-10; NFYB, p=2.37E-27; E2F4, p=7.44E-91; Mann-Whitney U, Fig. 5E], demonstrating that not only are ECGI highly enriched in TF motifs, they are much more likely to be bound by these TF proteins in chromatin relative to classical enhancers.

ECGI are Conserved as CpG Islands

The concentration of CpG-rich TF motifs in ECGI suggests that their CpG density may be fundamental to their ability to act as enhancers. We thus took two approaches to examine the conservation of ECGI across mammals and vertebrates: examining the frequency with which ECGI also met the UCSC criteria for CGI in other animals, and the phyloP [48] conservation scores for individual residues. This analysis revealed that ECGI are typically conserved as CGI throughout placental mammals (Fig. 6A), although not to the same extent as promoter CGI, and are rarely conserved as CGI in non-mammalian vertebrates (Fig. 6B). Notably, the mouse has significantly fewer CGI than most mammals (N=16026 vs. human N=28691 based on UCSC criteria), and they appear to have preferentially lost ECGI, rather than canonical promoter CGI (Fig. 6B). CpG dinucleotides in ECGI were likewise selectively conserved both among placental mammals (Fig. 6C, combined p=1.46E-65, Mann-Whitney U) and among vertebrates (Fig. 6D combined p=3.18E-27, Mann-Whitney U) relative to those in classical enhancers, while non-CpG residues exhibited similar mammalian conservation rates in ECGI and classical enhancers (combined p=0.101), and were even slightly less conserved across vertebrates than were those in classical enhancers (combined p=2.44E-5). Promoters are known to be more conserved than enhancers in general [49], and promoter CGI have even greater retention of CpG dinucleotides than do ECGI across mammals and vertebrates (Fig 6A,B), consistent with their higher likelihood of meeting CGI criteria in other animals. This may reflect the biological role of ECGI as enhancers that serve as adaptable accessories to genes, rather than as promoters critical to the integrity of specific genes.

We noted that relative to ECGI, Remnant CGI (orphan CGI with no evidence of enhancer activity in any of the analyzed cell lines) are less often conserved as CGI in placental mammals, prompting us to examine their chromatin state (Fig. 6E). We find that Remnants are strikingly depleted in CTCF binding, and instead exhibit heterochromatic H3K9me3 and H3K27me3 when compared to ECGI and classical enhancers active in HMEC. Consistent with their relative absence of CTCF binding[50], Remnant CGI are almost completely DNA methylated in embryonic stem cells (ENCODE), whereas most ECGI are protected from methylation (Fig. 6F). This distinction in conservation between ECGI and Remnants disappears in the more distantly related marsupials and monotremes, as well as in non-mammalian vertebrates (Fig. 6A,B), suggesting that ECGI may have diverged functionally early in the evolution of placental mammals, initiating the decay of Remnant CGI CpG density coinciding with the loss of selective pressure to remain unmethylated.

Active ECGI are Resistant to Methylation Changes in Cancer; Inactive ECGI are Prone

The selective conservation of CpG sites within ECGI relative to other enhancers suggests that, as seen at promoter CGI, there has been a selection against CpG methylation in these regions in the germline across evolutionary time, and that their hypomethylated status is critical to their function. To investigate possible mechanisms underlying the persistent hypomethylation of ECGI, we first examined the levels of 5-hydroxymethylation (5hmC), a known intermediate in passive and active DNA demethylation [51; 52], using a 5hmC Capture-Seq data set from IMR90 cells [53] (Fig. 7A). Strikingly, we find that ECGI from each cell line examined were uniquely marked by high levels of 5hmC, compared to both promoter CGI and classical enhancers from the same cell line, suggesting active turnover of DNA methylation. Promoter CGI are also suggested to be protected from DNA methylation by R-loops, RNA-DNA hybrids that form co-transcriptionally preferentially at G-skewed and GC-rich regions [2; 54; 55]. Interestingly, ECGI exhibit substantial enrichment for R-loops, approaching that of promoter CGI (Fig. 7B). In contrast, classical enhancers exhibit almost no detectable R-loop formation, consistent with the differences in DNA methylation and nascent transcription between these groups (see Fig. 1E and Fig. 2A).

The aberrant hypermethylation of typically unmethylated promoter CGI has been linked to tumor suppressor gene silencing [56]. Likewise, alterations in enhancer methylation state have been implicated in tumorigenesis, cancer progression, and metastasis [13; 57]. To investigate ECGI methylation in human cancers, we utilized 450k Infinium Methylation array data from 97 normal breast and 781 breast tumor samples (TCGA consortium) to compare the average methylation state of ECGI and classical enhancers active in HMEC versus those only active in other cell lines. We defined significant methylation changes as those with an FDR<.01 and an absolute change in β>0.2.

ECGI active in HMEC were unmethylated in normal tissue, and were resistant to methylation changes in primary breast tumors (3.4% hypermethylated, none hypomethylated). While most ECGI inactive in HMEC were also unmethylated, a substantial proportion are methylated in normal breast tissue (26.5% with average β >0.7). In cancer, these inactive ECGI were prone to hypermethylation, while few undergo hypomethylation (10.3% hypermethylated, 2.1% hypomethylated). These data suggest that persistent ECGI activity is necessary to repel aberrant DNA methylation (Fig. 7 C-F). In contrast, active classical HMEC enhancers exhibit a variable methylation pattern in normal breast tissue (median β=0.6, Fig. 7C-F), and are more prone to methylation changes than active ECGI (4.4% hypermethylated, 6.1% hypomethylated in active classical enhancers vs. 3.4% hypermethylated and 0% hypomethylated in active ECGI). As expected, classical enhancers inactive in HMEC tended to be more methylated in normal breast tissue than active ones (median β=0.72), but unlike inactive ECGI, inactive classical enhancers were more prone to hypomethylation (1.1% hypermethylated, 5.9% hypomethylated) in breast cancers. Thus, ECGI are highly resistant to methylation while active, but a subset of silent ECGI may gain methylation either normally during cell-type specification or aberrantly during carcinogenesis. In contrast, classical enhancers appear less resistant to DNA methylation than ECGI, even if active, which can lead to aberrant hypo- or hypermethylation in tumors. Together, these data show that the pervasive disparities between classical and ECGI in chromatin state, architecture, and conservation manifest as sweeping differences in DNA methylation dynamics during development and carcinogenesis.

Discussion

Research has largely focused on the canonical role of CGI as strong promoters, even though many 'orphan' CGI in the human genome exist far from any known transcript. At promoter CGI, the lack of DNA methylation ensures the preservation of CpG density, the binding of GC and CpG-binding transcription factors, and recruitment of active chromatin modifiers, ultimately creating a permissive environment for transcriptional initiation [1]. Indeed, several groups have also documented transcriptional initiation at orphan CGI, typically ascribing it to promoter function [27; 58]. However, Mendizabal et al. [18] recently noted that many CGI resemble enhancers in terms of chromatin state. Indeed, transcriptional initiation is also a key feature of enhancers, and recent work has shown that promoters and enhancers share a common epigenetic architecture, with the strongest enhancers resembling weak promoters in terms of their chromatin state [16]. Here, we demonstrate the extension of this promoter-enhancer relationship, establishing that most orphan CGI are in fact putative enhancers, or ECGI.

ECGI resemble classical enhancers in many ways, but possess the elevated GC content, hypomethylation, and CpG density that empower promoter CGI. Like at promoter CGI, these features license recruitment of TFs and chromatin modifiers, but result in enhancer-like H3K4me1 and unstable transcripts, rather than promoter-like H3K4me3 and stable transcripts. Just as promoter CGI are stronger and more euchromatic than other promoters, ECGI display higher eRNA production, histone acetylation, H3K4me3/me1 ratios, and functional ability to drive gene expression than do classical enhancers. The greater strength of ECGI is likely a direct reflection of their importance, and ECGI are also enriched in highly active 'super' enhancers that have been assigned pivotal roles in development, pluripotency, and oncogenesis by driving the expression of genes essential to the lineage and proliferation regulation [33; 59; 60].

Indeed, enhancer activity may be a feature of CGI in general, not just orphans. Critically, we find that many transcript-associated CGI also frequently overlap putative enhancers and exhibit strictly unstable transcripts. Additionally, in screens of enhancer activity even gene-associated CGI were more capable than other elements of enhancing expression.

This suggests a role as enhancers for thousands of additional genic CGI that lack a TSS. Embedded enhancer activity also may be a mechanism by which canonical CGI promoters achieve higher and broader gene expression across cell types than other promoters [1].

Nuclear architecture, the formation and location of DNA:DNA contacts, is central to gene regulation, enforcing chromatin boundaries and enabling enhancers to act on promoters [61]. ECGI display far more of these contacts that do classical enhancers, and they exhibit far greater enrichment of factors like CTCF, cohesin, and Brd4 that orchestrate these loops, suggesting they engage in more or tighter contacts. CTCF, which regulates recruitment of cohesin, likely prefers ECGI to classical enhancers because it is unable to bind methylated DNA [50; 62]. As enhancers must contact promoters to act, these loops are likely essential to ECGI activity, and recent work has linked aberrant nuclear disorganization to carcinogenesis [63]. Indeed, aberrant methylation of the PTSG2 promoter CGI in cancer cells abolishes CTCF/cohesin binding, silences the gene, and disrupts architecture across the locus [64].

CTCF is just one of many methylation-sensitive transcription factors that enhancers rely on for their activity [52]. As a whole, TF binding motifs have higher G+C content than the genome average [1], and are generally more common in both promoter CGI and ECGI than in classical enhancers. In chromatin, TFs with GC-rich and CpG-containing motifs accumulate to much greater levels at ECGI than at classical enhancers, and a number of these play critical roles in carcinogenesis. For example, SP1 overexpression occurs in many cancers and is linked to poor survival [65], ELK4 translocations can drive prostate cancer [66], and EGR1 regulates the survival of endocrine-resistant breast cancer cells [67]. Enrichment of these factors suggests that ECGI may play integral roles in carcinogenesis by mediating the ability of TFs to activate their target genes, or by acting as deep sinks to titrate TF pools, an effect which may be exacerbated by cancer-related methylation changes, perturbing TF binding and activity elsewhere [68].

The CpG density and hypomethylation that permit TF binding are linked evolutionarily by the mutagenicity of methylated cytosine. The conservation of promoter CGI stems from the specific conservation of CpG sites, rather than other nucleotides, suggesting that there is selective pressure to maintain CpG density and that this is the critical factor in their activity [69]. ECGI exhibit the same phenomena, but to a lesser extent than promoter CGI. This likely reflects the diminished H3K4me state of ECGI relative to promoter CGI, and the fact that enhancers evolve more rapidly than promoters [49]. Indeed, we find evidence for previously active ECGI in Remnant CGI; those CGI that are not transcript-associated and did not exhibit evidence of enhancer activity in any cell line. Unlike ECGI, Remnants are heavily methylated and heterochromatic, similar to the promoter CGI associated with pseudogenes, leading to their loss over time. Although some Remnants may represent the promoters of lost transcripts, many are likely decommissioned ECGI, given their distance from detectable pseudogenes. This finding highlights that a persistent function, and selection to remain unmethylated, is necessary to maintain CpG density, and suggests that ECGI have had important roles in mammalian evolution.

The conservation of CpG sites and hypomethylation suggest ECGI have mechanisms to repel DNA methylation. The TET enzymes, which catalyze the oxidation of 5-methyl cytosine residues, have been implicated in maintaining enhancer activity by preserving DNA hypomethylation [70]. Consistent with this idea, 5hmC is found at especially high levels in super enhancers [71], and DNA methylation preferentially accumulates in enhancer regions in cancers that experience loss of TET2 function [72]. We find that 5hmC is also heavily enriched at ECGI, compared to either promoter CGI or other enhancers. H3K4me3 and R-loops inhibit DNA methyltransferase recruitment at promoter CGI, but ECGI may lack full protection because of their lower H3K4me states and R-loop formation. Thus ECGI may experience higher rates of DNA methylation, eliciting a greater need to remove it. Alternatively, hydroxymethylation may serve a functional role at ECGI independent of its role as an intermediate in 5mC turnover, by either buffering the binding of 5hmC-sensitive TFs [73; 74] or by specifically recruiting 5hmC readers, several candidates of which have recently been identified [75].

It is well established that hypermethylation of promoter CGI often silences tumor suppressor genes during cancer progression, and that hypomethylation of intragenic CGI can unleash certain oncogenes like hTERT[76]. More recently, methylation changes at classical enhancers during carcinogenesis have been linked to altered activity and changes in gene expression [57; 77]. In fact, it has been suggested that enhancers exhibit more DNA methylation changes in cancer than other genomic compartments, and that these changes can modulate the expression of known oncogenes like KIT and ESR1 at a distance [77]. Here we show that ECGI inactive in HMEC cells (but active in another cell type) are especially prone to methylation changes, with more than 10% exhibiting significant hypermethylation in primary breast tumors. Bae et al. [17] recently suggested broad hypermethylation of enhancer-like CGI that lack TSS in cancer. However, they did not distinguish intragenic and other transcript-associated CGI in their analysis, which we find often contain stable transcripts (Supplemental Fig. 1B) making many likely promoters. Indeed, while we do find that more than half of ECGI active in HMECs have lost H3K27Ac (and presumably their activity) in MCF7 breast cancer cells, few of these become hypermethylated in primary breast tumors. Furthermore, we also find that dozens of ECGI inactive in HMEC acquire aberrant H3K27Ac in MCF7 cells (Supplemental Fig. 4). These findings suggest that, rather than indiscriminant hypermethylation, there is frequent decommissioning of active ECGI (often independent of methylation changes) as well as cryptic activation of inactive ECGI during oncogenesis. Given their massive activity and role in genome organization, even modest changes in the unique chromatin state of ECGI may impact cancer progression and survival by directly or indirectly perturbing gene expression.

Thus, we have identified that ECGI represent a novel class of enhancers, more powerful than classical enhancers by every measure and prone to aberrant DNA methylation changes in cancer. These findings point to a common evolutionary origin of CpG-rich promoters and enhancers, with enhancer CGI representing a subset of enhancers that possess many of the features of promoter CGI likely due to similar evolutionary pressures to maintain activity. This helps resolve the longstanding mystery of orphan CGI function and illuminates new aspects of enhancer and CpG island biology critical to understanding the chromatin dynamics that drive development and carcinogenesis.

Methods

CpG Island and Enhancer Annotation

CpG Islands as defined by UCSC (hg19) were annotated based on their relationship with GenCode (V25, hg19) transcripts, hierarchically by those 1) containing a protein-coding TSS, 2) contained within a protein-coding gene, 3) within 2kb of a protein-coding gene (perigenic), or within 2kb or overlapping: 4) a lncRNA, 5) other ncRNA (miRNA, rRNA, scRNA, snRNA, snoRNA, ribozyme, sRNA, antisense RNA, or scaRNA, 6) pseudogenes (including protein-coding or any other). Other CGI were considered to be 'orphan' CGI.

For enhancer definitions, cell lines examined as part of the ENCODE and Roadmap Epigenomics project that had available H3K4me1 and H3K27Ac ChIP-Seq called peaks or an HMM chromatin state map were used. ENCODE lines with peak defintitions were: A549, astrocytes, CD14+, skin fibroblasts, DND41, GM12878, H1ESC, HCT116, HeLa, HepG2, HMEC, HUVEC, K562, keratinocytes, lung fibroblasts, myotubes, osteoblasts, Panc1, and skeletal myoblasts, and with HMM definitions:GM12878, H1ESC, HepG2, HMEC, HSMM, HUVEC, K562, NHEK, and NHLF.

For Roadmap cells with the following EIDs were used for peaks and HMM defintion: E003, E004, E005, E006, E007, E008, E011, E012, E013,E014,E015,E016,E017,E019, E020, E021, E022, E026, E029, E032, E034, E037, E038, E039, E040, E041, E042, E043, E044, E045, E046, E047, E048, E049, E050, E055, E056, E058, E059, E061, E062, E063, E065, E066, E067, E068, E069, E071, E072, E073, E074, E075, E076, E078, E079, E080, E084, E085, E087, E089, E090, E091, E092, E093, E094, E095, E096, E097, E098, E099, E100, E101, E102, E103, E104, E105, E106, E108, E109, E111, E112, E113, E114, E115, E116, E117, E118, E119, E120, E121, E122, E123, E124, E125, E126, E127, E128, E129, or just for HMM: E001, E002, E009, E010, E018, E023, E024, E025, E027, E028, E030, E031, E033, E035, E036, E051, E052, E053, E054, E057, E070, E077, E081, E082, E083, E086, E088, E107, E110. For ENCODE HMMs, categories 4-7 (Strong and Weak Enhancers) were considered, and for Roadmap HMMs, categories 6,7, and 12 were used (Genic Enhancers, Enhancers, Bivalent Enhancers). For peak definitions, enhancers were defined as the overlap of H3K4me1 and H3K27ac peaks within each cell line. Orphan CGI that overlapped enhancers by both peak and HMM definition were used for ECGI analysis in the text. Given the lower resolution of HMM, where enhancers are used as a comparison, they represent regions of overlapping H3K4me1 and H3K27Ac peaks, that at least partially overlap an HMM enhancer.

As a negative control for the transcription stability analysis, five thousand random intergenic regions (at least 2kb from any Gencode transcript) of 500bp (for similarity to ECGI and enhancers) were chosen. Non-CGI promoters were defined as +/- 500bp from any protein-coding transcript in Gencode without a CGI within 2kb of the TSS.

For CGI from other animals, CGI tables were downloaded from UCSC, and lifted to hg19 using chains available from UCSC. Genomes used are as follows: bonobo (Pan paniscus, PanPan1), baboon (Papio anubis, PapAnu2), mouse (Mus musculus, mm9), cat (Felis catus, FelCat5), manatee (Trichechus manatus, TriMan1), opossum (Monodelphis domestica, MonDom5), platypus (Ornithorhynchus anatinus, OrnAna1), chicken (Gallus gallus, GalGal4), alligator (Alligator mississippiensis AllMis1), turtle (Chrysemys picta ChrPic1), xenopus (Xenopus tropicalis, XenTro3), coelacanth (Latimeria chalumnae, LatCha1), zebrafish (Danio rerio, DanRer10), lamprey (Petromyzon marinus, PetMar2).

ChIP-Seq and Chromatin Analyses

ENCODE ChIP-Seq datasets were downloaded as mapped BAM files. Tag densities for genomic features were determined using R/Bioconductor packages GenomicRanges, GenomicAlignments [78], and Rtracklayer[79], and visualized using ggplot2[80]. The Mann-Whitney U test is used for significance tests of differences in tag density. Accession numbers are as follows: HMEC H3K4me1 GSM733705, HMEC H3K4me2 GSM733654, HMEC H3K4me3 GSM733712, HMEC H3K27Ac GSM733660, HMEC H3K9Ac GSM733713, HMEC DNAse ENCSR000ENV, K562 H3K4me1 GSM733692 , K562 H3K27Ac GSM733656, K562 H3K9Ac GSM733778, K562 DNAse ENCFF000SVI, K562 H3K27me GSM733658, K562 H3K9me3 GSM733776, H1Esc H3K4me1 GSM733782 , H1Esc H3K27Ac GSM733718, H1Esc H3K9Ac GSM733773, H1Esc DNAse ENCFF658MCK, H1Esc H3K27me GSM733748, H1Esc H3K9me3 ENCFF769VJB. For data sets for which mapped files were unavailable (MCF7 GroSeq GSE27463), reads were mapped to hg19 using Bowtie2 [81] with default settings. For DRIP-Seq data (GSE57353), tag densities for the DRIP-Seq library (normal control fibroblasts) were normalized to that of the input library.

DNA Methylation Analysis

Processed (Tier 3) data for WGBS and Illumina Infinium 450k methylation array data were downloaded from TCGA (ID numbers in Supplementary Table 1). DNA methylation levels at each feature were calculated by averaging the β values (450k) or percent methylation (WGBS) for each CpG site or probe present in each genomic locus using the GenomicRanges R package [78]. The significance of hyper- and hypomethylation was determined by an FDR-corrected Wilcoxon rank sum test, with a 0.2 change in β used as an additional cutoff for significance.

Transcription Factor Binding Sites

JASPAR [47] database putative TFBS bed files for hg19 (~1.1 million binding sites for ~130 TFs) were downloaded. Motif GC content was determined by averaging the combined G+C likelihood of each residue in the motif. Motifs were considered to have a CpG site if they contained a site with at least a 50% likelihood of containing a C followed by a site with at least a 50% likelihood of containing a G. Odds ratios are the prevalence of a motif occurring within total kb contained by ECGI vs the prevalence in classical enhancers (see above).

Enhancer Screens

For lentiMPRA, a file containing each screened region and its score was downloaded from the supplement of the article [35]. In lentiMPRA, putative enhancer regions are cloned into a lentiviral GFP reporter vector that then integrates into the genome. Targeted RNA and DNA sequencing are performed, and the enhancer activity is determined by the ratio of RNA to DNA copy number. We present only the integrase-competent lentiMPRA analysis, but obtained similar results with the integrase-deficient library, although Inoue et al. note that the integrase-competent method is much more robust in detecting enhancers.

For mouse assays, FIREWACh[36] and CapStarr-Seq[37], files containing of regions in the input vs. captured libraries (and enhancer scores and categories for CapStarr-Seq) were downloaded from the supplementary materials of each publication. Both assays clone genomic regions into GFP enhancer-reporter vectors with subsequent transfection into cells, but FIREWACh clones putative enhancers upstream of the GFP promoter, which CapStarr-Seq clones the putative enhancers downstream of the reporter, enabling expression of the element in a GFP fusion transcript. FIREWACh relies on the isolation of GFP+ cells, identifying any element able to drive strong GFP reporter transcription, but without further quantification of strength. CapStarr-Seq utilizes targeted RNA sequencing to determine relative enrichment of each screened region (RNA copy number), relative to DNA copy number in the input library (cloned plasmids before transfection) to assign each element an enhancer score (fold change [FC] in input DNA copy number vs. RNA copy number in transfected cells). Based on this FC, the authors assigned each element to a category: Inactive (FC<1.5), Weak (FC 1.5-3), or Strong (FC>3).

Hi-C and ChIA-PET

For Hi-C, intrachromosomal combined contact matrices with scores for each annotated interaction in K562 cells were downloaded from GEO, as were TAD boundary locations in GM12878 (GSE63525). For ChIA-PET, bed files of annotated contacts were downloaded from ENCODE (CTCF in K562 cells ENCSR000CAC, POLR2A in MCF7 cells ENCSR000CAA).

Conservation

Phylop [48] conservation score files (bigWig) for mammalian and vertebrate genomes were downloaded from UCSC. For each feature, the average score for CpG residues and non-CpG residues was calculated independently.

Acknowledgements

We would like to acknowledge Benjamin Barwick for thoughtful editing of the manuscript. We would also like to acknowledge the patients, donors, and scientists of the TCGA, ENCODE, and Roadmap Epigenome projects, without which this project would not be possible.

References

1. Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. Genes & development 25(10):1010-1022.
2. Kellner WA, Bell JS, Vertino PM. 2015. GC skew defines distinct RNA polymerase pause sites in CpG island promoters. Genome research 25(11):1600-1609.
3. Zhang Y, Ng H-H, Erdjument-Bromage H, Tempst P, Bird A, Reinberg D. 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes & development 13(15):1924-1935.
4. Baylln SB, Herman JG, Graff JR, Vertino PM, Issa J-P. 1997. Alterations in DNA methylation: a fundamental aspect of neoplasia. Advances in cancer research 72:141-196.
5. Conway KE, McConnell BB, Bowring CE, Donald CD, Warren ST, Vertino PM. 2000. TMS1, a novel proapoptotic caspase recruitment domain protein, is a target of methylation-induced gene silencing in human breast cancers. Cancer Research 60(22):6236-6242.
6. Issa J, Vertino PM, Boehm CD, Newsham IF, Baylin SB. 1996. Switch from monoallelic to biallelic human IGF2 promoter methylation during aging and carcinogenesis. Proceedings of the National Academy of Sciences 93(21):11757-11762.
7. Esteller M, Hamilton SR, Burger PC, Baylin SB, Herman JG. 1999. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. Cancer Research 59(4):793-797.
8. Grady WM, Willis J, Guilford PJ, Dunbier AK, Toro TT, Lynch H, Wiesner G, Ferguson K, Eng C, Park J-G. 2000. Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. Nature genetics 26(1):16-17.
9. Kane MF, Loda M, Gaida GM, Lipman J, Mishra R, Goldman H, Jessup JM, Kolodner R. 1997. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. Cancer Research 57(5):808-811.
10. Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99(3):247-257.
11. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130(1):77-88.
12. Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic acids research 8(7):1499-1504.
13. Aran D, Hellman A. 2013. DNA methylation of transcriptional enhancers and cancer predisposition. Cell 154(1):11-13.
14. Barwick BG, Scharer CD, Bally AP, Boss JM. 2016. Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation. Nature Immunology 17(10):1216-1225.
15. Bell JS, Kagey JD, Barwick BG, Dwivedi B, McCabe MT, Kowalski J, Vertino PM. 2016. Factors affecting the persistence of drug-induced reprogramming of the cancer methylome. Epigenetics 11(4):273-287.
16. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics 46(12):1311-1320.
17. Bae MG, Kim JY, Choi JK. 2016. Frequent hypermethylation of orphan CpG islands with enhancer activity in cancer. BMC Medical Genomics 9(1):38.
18. Mendizabal I, Soojin VY. 2015. Whole-genome bisulfite sequencing maps from multiple human

tissues reveal novel CpG islands associated with tissue-specific regulation. Human molecular genetics:ddv449.

19. Bock C, Walter J, Paulsen M, Lengauer T. 2007. CpG island mapping by epigenome prediction. PLoS Comput Biol 3(6):e110.
20. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research 22(9):1775-1789.
21. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22(9):1760-1774.
22. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 107(50):21931-21936.
23. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470(7333):279-283.
24. Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.
25. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ. 2015. Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317-330.
26. Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. Nature methods 9(3):215-216.
27. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D’Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466(7303):253-257.
28. Network CGA. 2012. Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61-70.
29. Barrero MJ, Sese B, Kuebler B, Bilic J, Boue S, Martí M, Belmonte JCI. 2013. Macrohistone variants preserve cell identity by preventing the gain of H3K4me2 during reprogramming to pluripotency. Cell reports 3(4):1005-1011.
30. Fang R, Barbera AJ, Xu Y, Rutenberg M, Leonor T, Bi Q, Lan F, Mei P, Yuan G-C, Lian C. 2010. Human LSD2/KDM1b/AOF1 regulates gene transcription by modulating intragenic H3K4me2 methylation. Molecular cell 39(2):222-233.
31. Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322(5909):1845-1848.
32. Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A. 2015. Identification of active transcriptional regulatory elements from GRO-seq data. Nature methods 12(5):433-438.
33. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153(2):307-319.
34. Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, Wang F, Su J, Wu Q, Liu H. 2016. SEA: a super-enhancer archive. Nucleic acids research 44(D1):D172-D179.
35. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2016. A systematic comparison reveals substantial differences in chromosomal versus episomal

encoding of enhancer activity. bioRxiv:061606.

36. Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, Xi X, Basilico C, Brown S, Bonneau R. 2014. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nature methods 11(5):559-565.
37. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LT, Fernandez N, Ballester B, Andrau JC, Spicuglia S. 2015. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nature communications 6.
38. Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D. 2009. Transcription factors mediate long-range enhancer–promoter interactions. Proceedings of the National Academy of Sciences 106(48):20222-20227.
39. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. Nature genetics 43(7):630-638.
40. Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics 12(4):283-293.
41. Kanno T, Kanno Y, LeRoy G, Campos E, Sun H-W, Brooks SR, Vahedi G, Heightman TD, Garcia BA, Reinberg D. 2014. BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. Nature structural & molecular biology 21(12):1047-1057.
42. Liu W, Ma Q, Wong K, Li W, Ohgi K, Zhang J, Aggarwal AK, Rosenfeld MG. 2013. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. Cell 155(7):1581-1595.
43. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665-1680.
44. Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, Goh Y, Fullwood MJ, Sung W-K, Ruan X. 2012. ChIA-PET analysis of transcriptional chromatin interactions. Methods 58(3):289-299.
45. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398):376-380.
46. Barozzi I, Simonatto M, Bonifacio S, Yang L, Rohs R, Ghisletti S, Natoli G. 2014. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Molecular cell 54(5):844-857.
47. Mathelier A, Fornes O, Arenillas DJ, Chen C-y, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R. 2015. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic acids research:gkv1176.
48. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research 20(1):110-121.
49. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ. 2015. Enhancer evolution across 20 mammalian species. Cell 160(3):554-566.
50. Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature 405(6785):486-489.
51. Guo JU, Su Y, Zhong C, Ming G-l, Song H. 2011. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell 145(3):423-434.
52. Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, Zhang X, Cheng X. 2012. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic acids research:gks155.

53. Wang T, Wu H, Li Y, Szulwach KE, Lin L, Li X, Chen I-P, Goldlust IS, Chamberlain SJ, Dodd A. 2013. Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency. Nature cell biology 15(6):700-711.
54. Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Molecular cell 45(6):814-825.
55. Lim YW, Sanz LA, Xu X, Hartono SR, Chédin F. 2015. Genome-wide DNA hypomethylation and RNA: DNA hybrid accumulation in Aicardi–Goutieres syndrome. Elife 4:e08007.
56. McCabe MT, Brandes JC, Vertino PM. 2009. Cancer DNA methylation: molecular mechanisms and clinical implications. Clinical Cancer Research 15(12):3927-3937.
57. Aran D, Sabato S, Hellman A. 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome biology 14(3):1.
58. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet 6(9):e1001134.
59. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. Cell 155(4):934-947.
60. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. 2015. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Molecular cell 58(2):362-370.
61. Van Bortle K, Corces VG. 2013. The role of chromatin insulators in nuclear architecture and genome function. Current opinion in genetics & development 23(2):212-218.
62. Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. Cell 137(7):1194-1211.
63. Taberlay PC, Achinger-Kawecka J, Lun AT, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC. 2016. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome research 26(6):719-731.
64. Kang J, Song S, Yun J, Jeon M, Kim H, Han S, Kim T. 2015. Disruption of CTCF/cohesin-mediated high-order chromatin structures by DNA methylation downregulates PTGS2 expression. Oncogene 34(45):5677-5684.
65. Beishline K, Azizkhan‐Clifford J. 2015. Sp1 and the ‘hallmarks of cancer’. FEBS journal 282(2):224-258.
66. Rickman DS, Pflueger D, Moss B, VanDoren VE, Chen CX, de la Taille A, Kuefer R, Tewari AK, Setlur SR, Demichelis F. 2009. SLC45A3-ELK4 is a novel and frequent erythroblast transformation–specific fusion transcript in prostate cancer. Cancer Research 69(7):2734-2738.
67. Shajahan-Haq AN, Cheema A, Jin L, Boca S, Gusev Y, Bhuvaneshwar K, Demas D, Raghavan K, Madhavan S, Clarke R. 2016. Integration of transcriptomic and metabolomic data reveals a central role for EGR1 in regulating survival and cellular metabolism in endocrine-resistant breast cancer. Cancer Research 76(14 Supplement):1508-1508.
68. Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. 2014. The transcription factor titration effect dictates level of gene expression. Cell 156(6):1312-1323.
69. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nature genetics 38(6):626-635.
70. Hon GC, Song C-X, Du T, Jin F, Selvaraj S, Lee AY, Yen C-a, Ye Z, Mao S-Q, Wang B-A. 2014. 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. Molecular cell 56(2):286-297.
71. Johnson KC, Houseman EA, King JE, von Herrmann KM, Fadul CE, Christensen BC. 2016. 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in

glioblastoma patients. Nature communications 7:13177.

72. Rasmussen KD, Jia G, Johansen JV, Pedersen MT, Rapin N, Bagger FO, Porse BT, Bernard OA, Christensen J, Helin K. 2015. Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. Genes & development 29(9):910-922.
73. Liu Y, Zhang X, Blumenthal RM, Cheng X. 2013. A common mode of recognition for methylated CpG. Trends in biochemical sciences 38(4):177-183.
74. Wang D, Hashimoto H, Zhang X, Barwick BG, Lonial S, Boise LH, Vertino PM, Cheng X. 2016. MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. Nucleic acids research:gkw1184.
75. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PW, Bauer C, Münzel M, Wagner M, Müller M, Khan F. 2013. Dynamic readers for 5-(hydroxy) methylcytosine and its oxidized derivatives. Cell 152(5):1146-1159.
76. Nagarajan RP, Zhang B, Bell RJ, Johnson BE, Olshen AB, Sundaram V, Li D, Graham AE, Diaz A, Fouse SD. 2014. Recurrent epimutations activate gene body promoters in primary glioblastoma. Genome research 24(5):761-774.
77. Bell RE, Golan T, Malcov H, Amar D, Salamon A, Liron T, Sheinboim D, Gelfman S, Gabet Y, Shamir R. 2016. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. Genome research:gr. 197194.115.
78. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol 9(8):e1003118.
79. Lawrence M, Gentleman R, Carey V. 2009. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25(14):1841-1842.
80. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer Science & Business Media.
81. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9(4):357-359.

Figure Legends

Figure 1: Most orphan CpG islands exhibit an enhancer chromatin state

A) CpG Islands were categorized hierarchically by distance to the nearest Gencode annotated gene. Shown is the number of CGI that overlap the TSS of a protein-coding gene (TSS), those that overlap a protein coding gene but not the TSS (intragenic), those that lie in proximity (+/- 2kb) of a protein coding gene (perigenic), or those that are within 2kb of, or overlap, a long non-coding RNA (lncRNA), other non-coding RNAs (ncRNA), or a pseudogene. CGI more than 2kb from any of these gene classes were considered orphan CGI (see Methods). B) Enhancers were defined either as the overlap of H3K4me1/H3K27Ac peaks or by an HMM chromatin state as annotated in over 100 cell lines (see Methods). The histograms represent the number of cell lines in which orphan CGIs overlap an enhancer by each definition. C) Distribution of orphan CGI meeting one or both enhancer definitions in at least one cell line. D) Stable or Unstable transcript pairs as defined in K562 cells [15] were intersected with promoter CGI, the TSS (+/- 500bp) of other coding genes that not within 2.5kb of a CGI, ECGI, and classical enhancers active in K562 cells. Shown is the fraction of each set of genomic regions that overlap stable, unstable or mixed transcript pairs. E) Distribution of the average DNA methylation in WGBS data from normal breast tissue (TCGA) across promoter CGI (those overlapping a coding TSS), ECGI active in HMEC cells, and classical enhancers active in HMEC cells (those orphan CGI or other regions meeting both the H3K27Ac/H3K4me1 peak and HMM enhancer definition). F) Density of the G+C content (%GC) and CpG content (Observed/Expected) among promoter CGI, versus ECGI and classical enhancers active in HMEC cells. G-N) Analysis of H3K27Ac or H3K4me1/2/3 at promoter CGI, HMEC ECGI, and HMEC classical enhancers. G-J, Distribution of the density (reads/kb) for the indicated chromatin mark among genomic loci in each class. Line indicates median, boxes are the first and third quartiles and whiskers represent the highest and lowest values within 1.5 times the inter-quartile range. K-N relative tag densities for the indicated chromatin mark was determined in 10bp bins for +/- 2.5kb from the center of each genomic feature class as determined from ChIP-seq data from HMEC cells (ENCODE). O) Distribution of the ratio of H3K4me3 to H3K4me1 tag densities across genomic loci in each class.

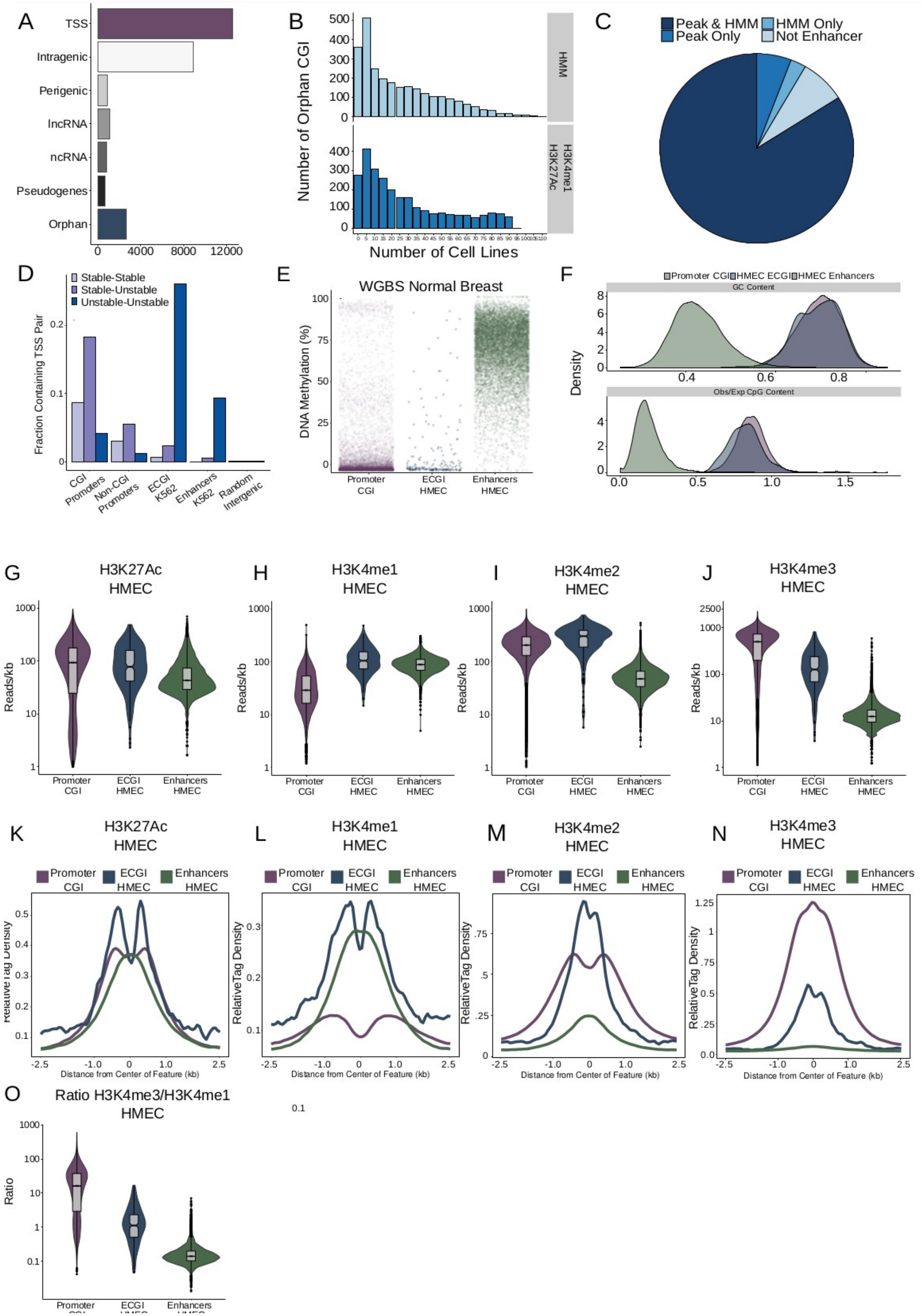

A
TSS
Intragenic
Perigenic
lncRNA
ncRNA
Pseudogenes
Orphan
0
4000
8000
12000
B
Number of Orphan CGI
HMM
H3K4me1
H3K27Ac
Number of Cell Lines
C
Peak & HMM
HMM Only
Peak Only
Not Enhancer
D
Stable-Stable
Stable-Unstable
Unstable-Unstable
Fraction Containing TSS Pair
CGI Promoters
Non-CGI Promoters
ECGI K562
Enhancers K562
Random Intergenic
E
WGBS Normal Breast
DNA Methylation (%)
Promoter CGI
ECGI HMEC
Enhancers HMEC
F
Promoter CGI
HMEC ECGI
HMEC Enhancers
GC Content
Obs/Exp CpG Content
Density
G
H3K27Ac HMEC
H
H3K4me1 HMEC
I
H3K4me2 HMEC
J
H3K4me3 HMEC
Reads/kb
Promoter CGI
ECGI HMEC
Enhancers HMEC
K
H3K27Ac HMEC
L
H3K4me1 HMEC
M
H3K4me2 HMEC
N
H3K4me3 HMEC
Promoter CGI
ECGI HMEC
Enhancers HMEC
Relative Tag Density
Distance from Center of Feature (kb)
O
Ratio H3K4me3/H3K4me1 HMEC
Ratio
Promoter CGI
ECGI HMEC
Enhancers HMEC
0.1

Figure 2: ECGI are stronger than CpG-poor enhancers

A-D) Nascent transcription and chromatin features associated with open/ active chromatin at ECGI and classical enhancers. Shown is the distribution of tag densities (reads/kb) of nascent transcription (GroSeq, MCF7 cells), active chromatin features (H3K27Ac, H3K9Ac ChIP-seq), or open chromatin (DNAse-seq; HMEC cells) across genomic loci in each enhancer class. E) Promoter CGI, ECGI, and classical enhancers active in the indicated cell type were overlapped with genomic regions called as super enhancers as defined by the Super Enhancer Archive (SEA). Shown is the fraction of genomic loci in each class overlapping any super enhancer in SEA. F) Distribution of enhancer activity scores, defined by the ratio of GFP reporter mRNA to DNA copy number in lentiMPRA (see Methods).

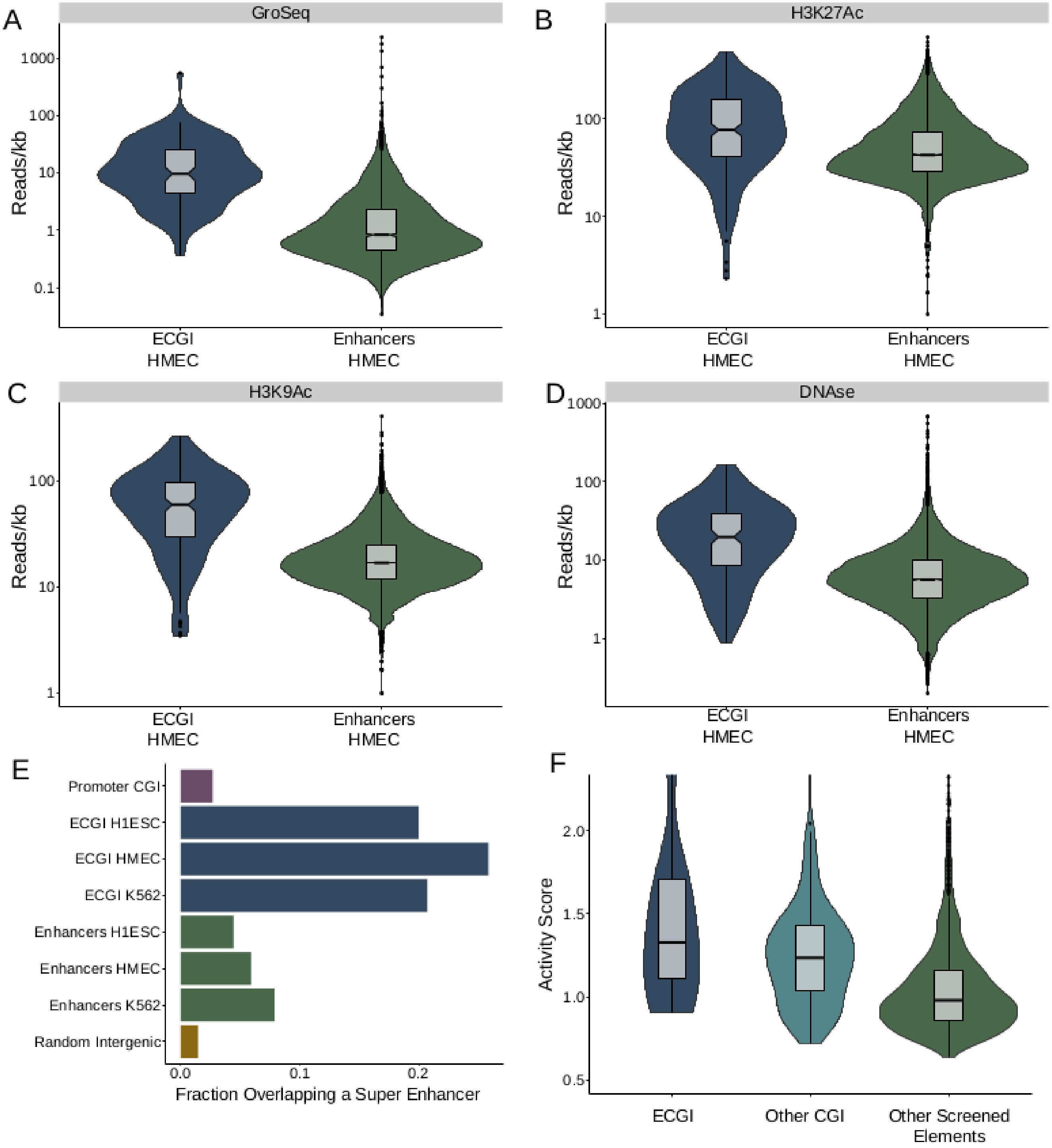

A
GroSeq
Reads/kb
1000
100
10
1
0.1
ECGI
HMEC
Enhancers
HMEC
B
H3K27Ac
Reads/kb
100
10
1
ECGI
HMEC
Enhancers
HMEC
C
H3K9Ac
Reads/kb
100
10
1
ECGI
HMEC
Enhancers
HMEC
D
DNAse
Reads/kb
1000
100
10
1
ECGI
HMEC
Enhancers
HMEC
E
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enhancers H1ESC
Enhancers HMEC
Enhancers K562
Random Intergenic
0.0
0.1
0.2
Fraction Overlapping a Super Enhancer
F
Activity Score
2.0
1.5
1.0
0.5
ECGI
Other CGI
Other Screened
Elements

Figure 3: ECGI are more widely active than CpG-poor enhancers

A) Distribution of thepercent of analyzed cell lines (N=120) in which the loci in each class exhibit enhancer activity (overlap an annotated H3K4me1/H3K27Ac peak in that cell type). B-G) Distribution of the ChIP-seq or DNAse-seq tag density (reads/kb) for the indicated chromatin feature as measured in HMEC cells (ENCODE) among ECGI or classical enhancers active in the indicated cell type.

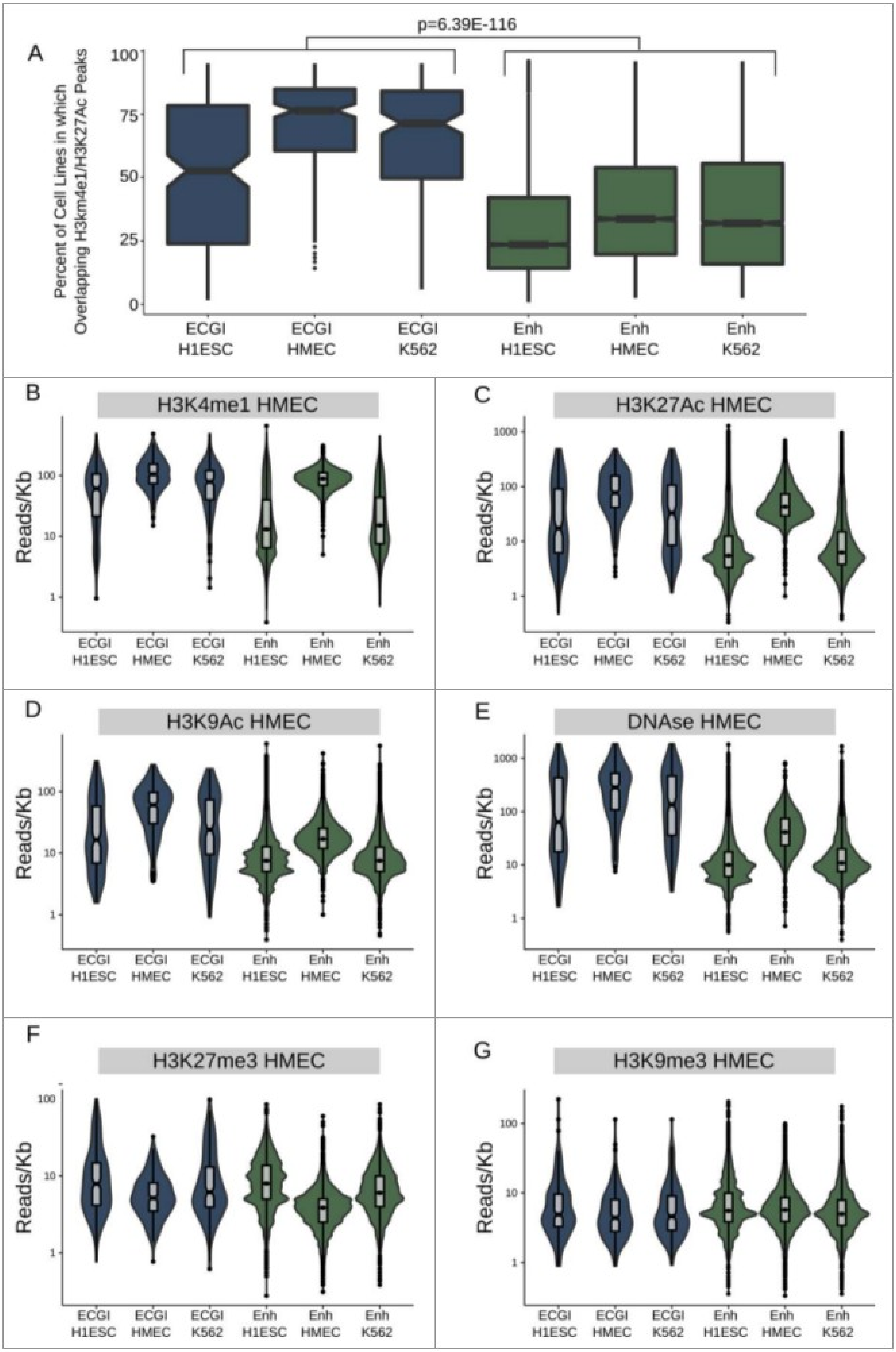

A
p=6.39E-116
Percent of Cell Lines in which Overlapping H3km4e1/H3K27Ac Peaks
100
75
50
25
0
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
B
H3K4me1 HMEC
C
H3K27Ac HMEC
D
H3K9Ac HMEC
E
DNAse HMEC
F
H3K27me3 HMEC
G
H3K9me3 HMEC
Reads/Kb
1000
100
10
1

Figure 4: ECGI are Hubs for Genomic Contacts

A) Distribution of the average CTCF (from H1ESC), cohesin/Rad21(from MCF7 cells), or BRD4 (from MM.1S) ChIP-seq tag densities (reads/kb) among promoter CGI versus ECGI or classical enhancers active in the indicated cell type. Genomic loci defined as in Figure 1. B) Annotated intrachromosomal contacts as defined by Rao et al. [42] were extracted from K562 cell Hi-C data, overlapped with the genomic loci in each class, and the average strength of all contacts per locus was determined. Shown is the distribution of the mean intrachromosomal contact strength among promoter CGI versus ECGI or classical enhancers active in the indicated cell type. C) The fraction of loci in each genomic class that overlap an annotated contact as determined by ChIA-Pet of CTCF in K562 cells or of RNA polymerase (POLR2A) in MCF7 cells (ENCODE). D) Median distance from TAD boundaries among genomic loci in each class as called in GM12878 cell Hi-C data [42].

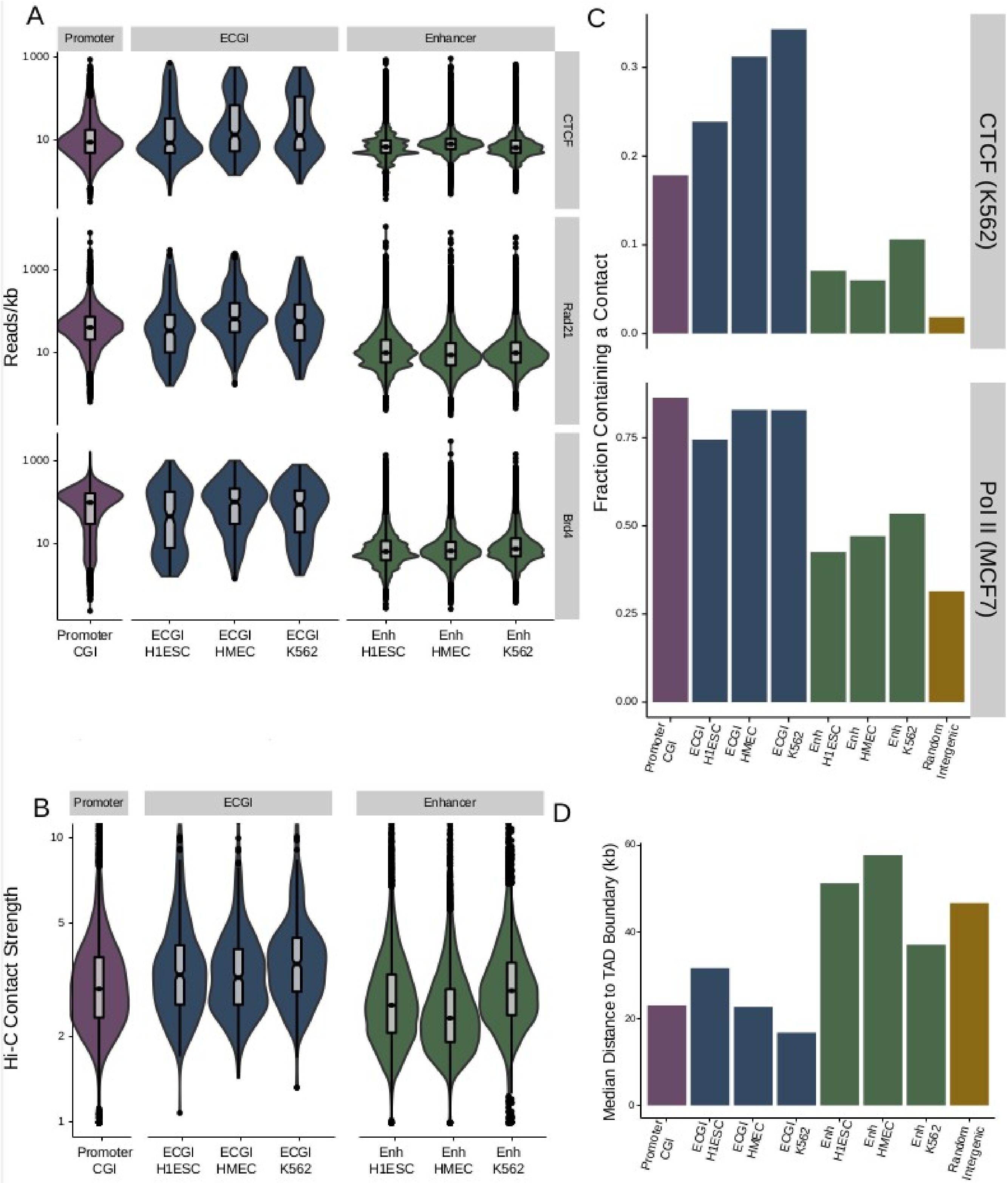

A
Promoter
ECGI
Enhancer
CTCF
Rad21
Brd4
Reads/kb
1000
10
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
C
CTCF (K562)
Pol II (MCF7)
Fraction Containing a Contact
0.0
0.1
0.2
0.3
0.00
0.25
0.50
0.75
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
Random Intergenic
B
Promoter
ECGI
Enhancer
Hi-C Contact Strength
1
2
5
10
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
D
Median Distance to TAD Boundary (kb)
0
20
40
60
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
Random Intergenic

Figure 5: ECGI are enriched in transcription factor binding

A) Distribution of the density (per kb) of transcription factor binding sites (Jaspar database) among promoter CGI versus ECGI or classical enhancers active in the indicated cell type. B) Relative enrichment of each transcription factor binding motif in the cumulative pool of ECGI called in HMEC, K562, and H1ESCs relative to that in all classical enhancers from the same cell lines. Shown is the Odds Ratio of enrichment of each motif among ECGI versus that of classical enhancers, +/- 95% confidence interval for each motif for which the ratio was >1 ($p<0.05$, Fisher's exact) . The GC content of each motif is indicated by the blue shaded bar color, and the error bar color indicates whether the motif contains a CpG (red with CpG; black without CpG). C) Relationship between GC content of enriched motifs vs. OR of enrichment. Shown is the linear regression of the relationship, with shadows representing the 95% confidence interval. D) Distribution of Odds Ratios of enrichment for those enriched motifs that contain or do not contain a CpG site. E) Distribution of the ChIP-seq tag densities (reads/kb) for representative transcription factors whose motifs are enriched in ECGI among genomic loci classified as ECGI or classical enhancers active in K562 cells.

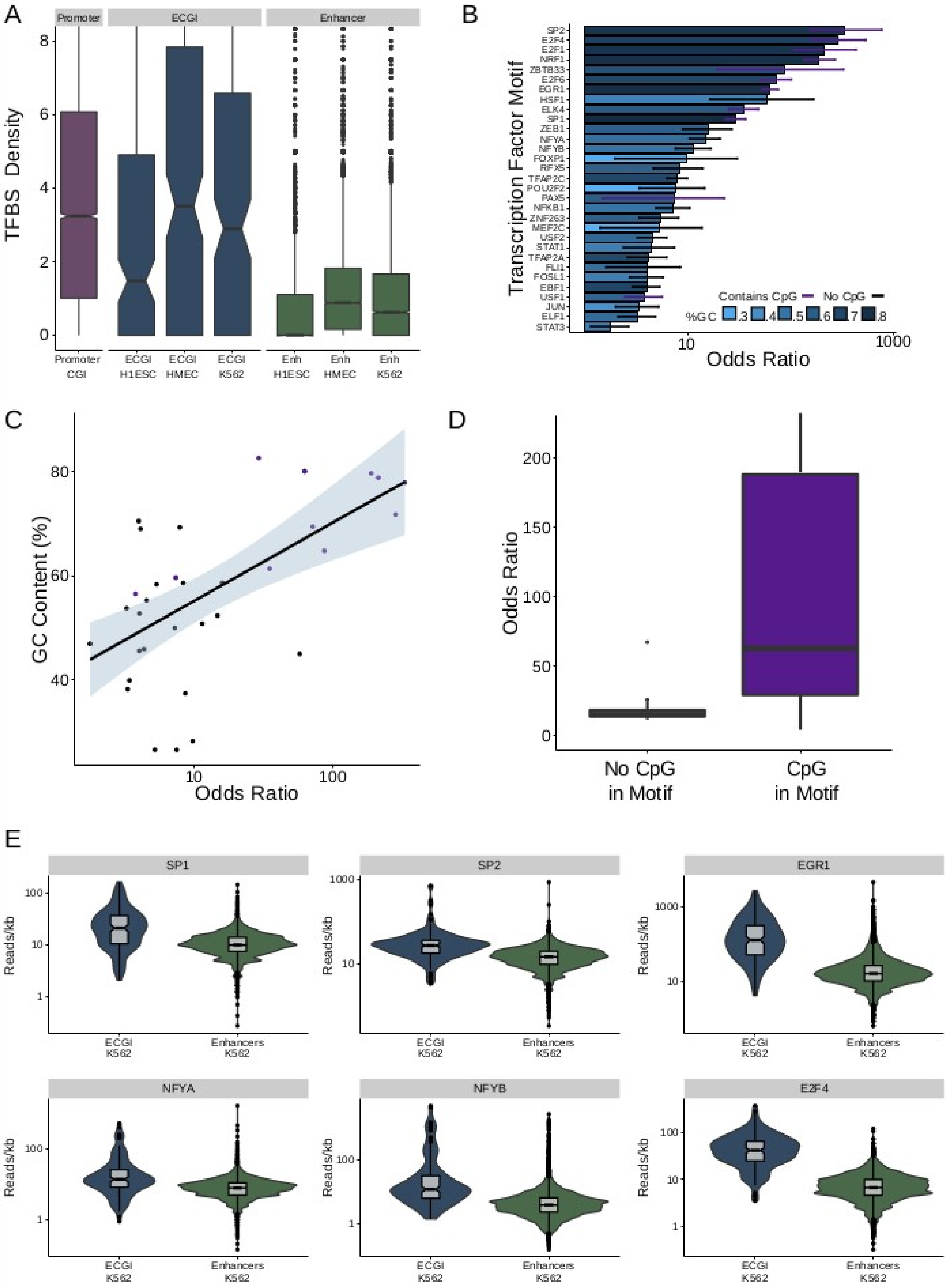
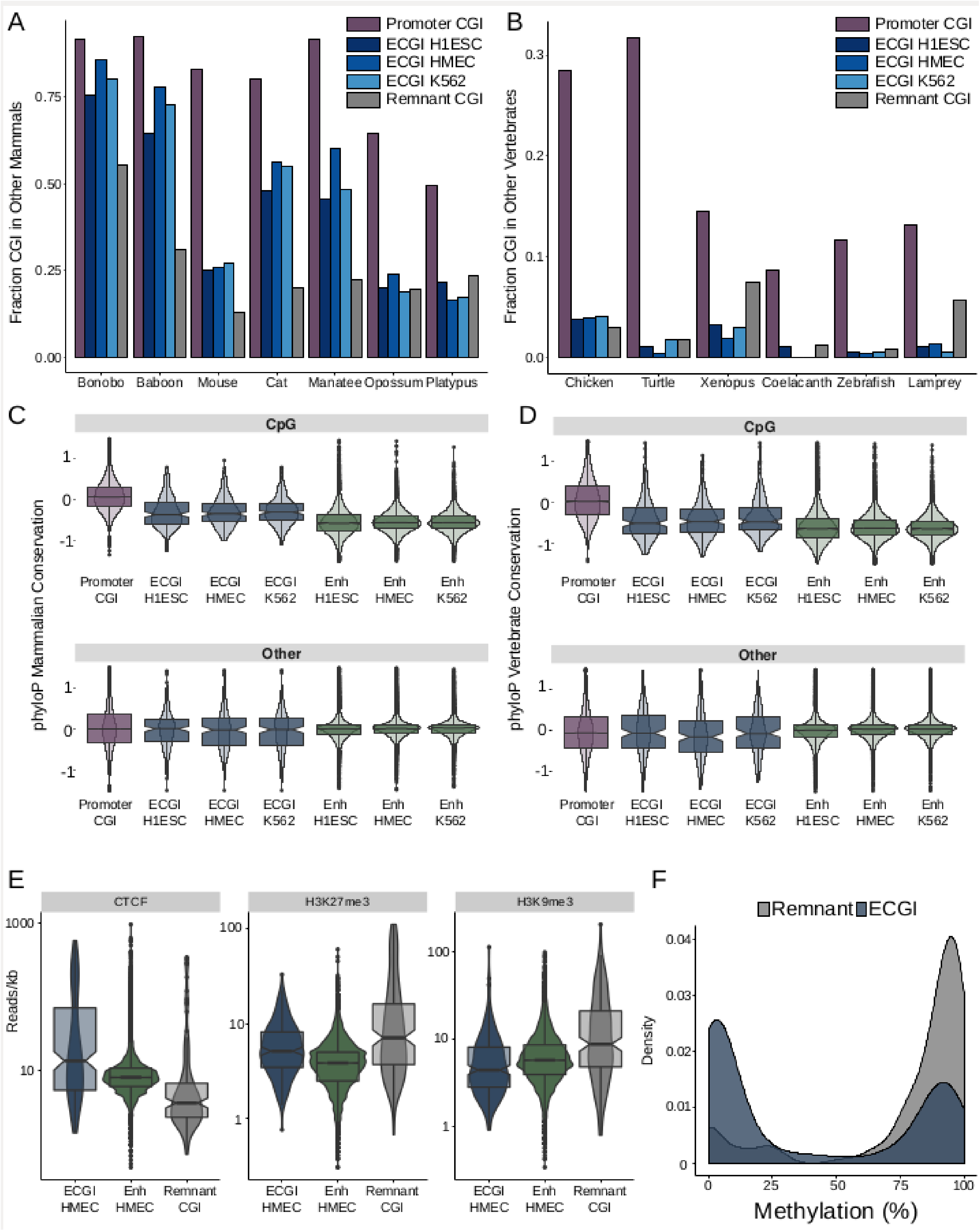

A
Promoter
ECGI
Enhancer
TFBS Density
0
2
4
6
8
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
B
Transcription Factor Motif
SP2
E2F4
E2F1
NRF1
ZBTB33
E2F6
EGR1
HSF1
ELK4
SP1
ZEB1
NFYA
NFYB
FOXP1
RFX5
TFAP2C
POU2F2
PAX5
NFKB1
ZNF263
MEF2C
USF2
STAT1
TFAP2A
FLI1
FOSL1
EBF1
USF1
JUN
ELF1
STAT3
Contains CpG
No CpG
%GC .3 .4 .5 .6 .7 .8
10
1000
Odds Ratio
C
GC Content (%)
40
60
80
10
100
Odds Ratio
D
Odds Ratio
0
50
100
150
200
No CpG in Motif
CpG in Motif
E
SP1
Reads/kb
1
10
100
ECGI K562
Enhancers K562
SP2
Reads/kb
10
1000
ECGI K562
Enhancers K562
EGR1
Reads/kb
10
1000
ECGI K562
Enhancers K562
NFYA
Reads/kb
1
100
ECGI K562
Enhancers K562
NFYB
Reads/kb
1
100
ECGI K562
Enhancers K562
E2F4
Reads/kb
1
10
100
ECGI K562
Enhancers K562

Figure 6: ECGI CpG density is highly conserved in mammals

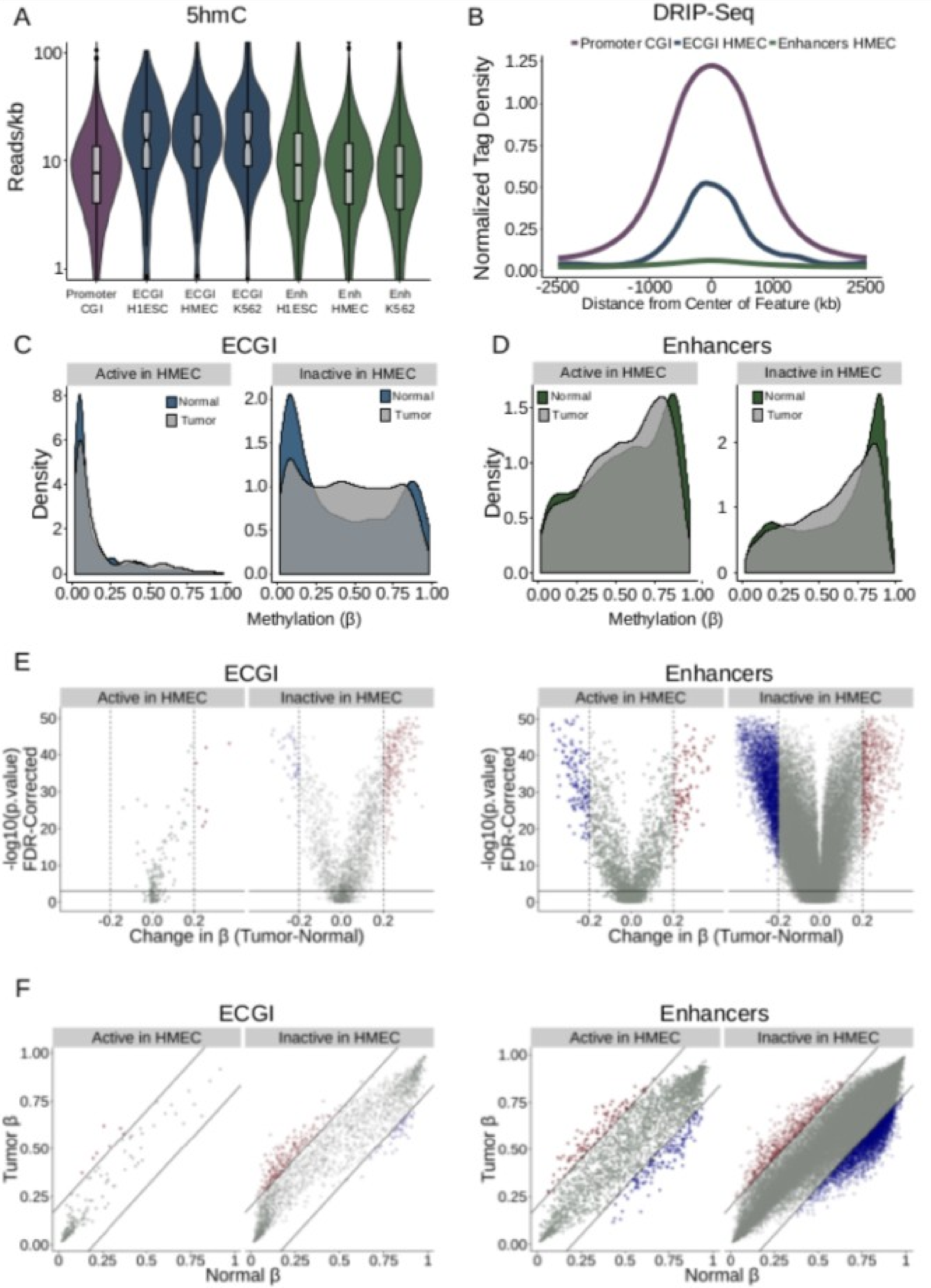
Genomic coordinates annotated as CGI from each species indicated (UCSC) were lifted over to the human genome (hg19). Shown is the fraction of human CGI in each class that overlapped a CGI in the indicated mammals (A) or other vertebrates (B). Distribution of the average placental mammal (C) or vertebrate (D) phyloP score for CpG dinucleotides (*top*) or other residues (*bottom*) among human promoter CGI, ECGI, or classical enhancers active in the indicated cell type. E) Distribution of average ChIP-Seq tag densities (read/kb) for the indicated chromatin feature (CTCF, H3K27me3, H3K9me3) among ECGI, classical enhancers and Remnant CGI as defined in HMEC cells. F) Density of the mean DNA methylation level for ECGI or Remnant CGI as determined from H1ESC whole genome bisulfite sequencing.



A
Fraction CGI in Other Mammals
0.00
0.25
0.50
0.75
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Remnant CGI
Bonobo
Baboon
Mouse
Cat
Manatee
Opossum
Platypus
B
Fraction CGI in Other Vertebrates
0.0
0.1
0.2
0.3
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Remnant CGI
Chicken
Turtle
Xenopus
Coelacanth
Zebrafish
Lamprey
C
phyloP Mammalian Conservation
CpG
Other
1
0
-1
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
D
phyloP Vertebrate Conservation
CpG
Other
1
0
-1
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
E
Reads/kb
CTCF
H3K27me3
H3K9me3
1000
100
10
1
ECGI HMEC
Enh HMEC
Remnant CGI
F
Remnant
ECGI
Density
0
0.01
0.02
0.03
0.04
0
25
50
75
100
Methylation (%)

Figure 7: Active ECGI are resilient to aberrant methylation in cancer

A) Distribution in tag density (reads/kb) of 5-hydroxymethylcytosine (5hmC) among promoter CGI versus ECGI or classical enhancers active in the indicated cell type (5hmC Capture-seq data; IMR90 cells). B) Mean normalized DRIP-seq (primary fibroblasts) tag densities in 10bp bins for +/- 2.5kb from the center of promoter CGI versus ECGI and classical enhancers active in HMEC. C, D) Density of the average methylation level (β) per feature as determined by 450k Methylation Array in normal breast tissue (n=97) or primary breast tumors (n=781, TCGA). ECGI active in HMEC vs. inactive in HMEC are those CGI that overlap H3K27Ac/H3K4me1 peak and HMM enhancer definitions in HMEC versus those called active in at least one other cell line assayed by the same criteria but absent from HMEC. Classical enhancers active versus inactive in HMEC were similarly defined but do not overlap a CGI. E) Volcano plot showing the change in the average DNA methylation (β value) per genomic feature among normal or breast tumor samples. Blue features are those ECGI/Enhancers significantly hypomethylated (FDR<.01 and change in β < -0.2), and red features are those signficantly hypermethylated (FDR<.01 and change in β > 0.2). Active and inactive ECGI versus classical enhancers are as defined in panels C,D. F) Relationship between the average DNA methylation (β value) for each ECGI/ Enhancer between normal and breast tumor samples. Lines represent an average change in β of 0.2. Active and inactive ECGI versus classical enhancers are as defined in panel C,D.


A
5hmC
Reads/kb
100
10
1
Promoter CGI
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
B
DRIP-Seq
Promoter CGI
ECGI HMEC
Enhancers HMEC
Normalized Tag Density
1.25
1.00
0.75
0.50
0.25
0.00
-2500
-1000
0
1000
2500
Distance from Center of Feature (kb)
C
ECGI
Active in HMEC
Inactive in HMEC
Normal
Tumor
Density
8
6
4
2
0
2.0
1.5
1.0
0.5
0.0
0.00 0.25 0.50 0.75 1.00
Methylation (β)
D
Enhancers
Active in HMEC
Inactive in HMEC
Normal
Tumor
Density
1.5
1.0
0.5
0.0
2
1
0
0.00 0.25 0.50 0.75 1.00
Methylation (β)
E
ECGI
Enhancers
Active in HMEC
Inactive in HMEC
-log10(p.value) FDR-Corrected
50
40
30
20
10
0
-0.2
0.0
0.2
Change in β (Tumor-Normal)
F
ECGI
Enhancers
Active in HMEC
Inactive in HMEC
Tumor β
1.00
0.75
0.50
0.25
0.00
0 0.25 0.5 0.75 1
Normal β

Supplemental Data

Supplemental Results :

As an additional method to assess the relationship between CpG conservation and ECGI enhancer activity, we analyzed mouse datasets generated with two functional enhancer screening protocols, FIREWACh (1) and CapStarr-Seq (2, 3). FIREWACh identifies putative regulatory elements by randomly cloning nucleosome-free genomic regions upstream of a GFP reporter, followed by transfection and FACS isolation of GFP+ cells, and targeted DNA sequencing to identify promoters and enhancers. CapStarr-Seq takes a similar approach, but clones putative regulatory elements downstream of the GFP reporter to specifically identify enhancers followed by targeted RNA-Seq of plasmid transcripts to provide a more quantitative measure of activity. We analyzed published mouse FIREWACh and CapStarr-seq experiments, comparing screened elements (FIREWACh N=84,240, CapStarr-seq N=7,542) that, upon lift over to the human genome, were either a) 'conserved ECGI' (overlapping a human ECGI and a mouse CGI; FIREWACh N=193,CapStarr-seq N=17), b) 'nonconserved ECGI' overlapping a human ECGI but not a mouse CGI (FIREWACh N=170, CapStarr-seq N=26), c) 'other mouse CGI' (overlapping a mouse CGI, but not a human ECGI, but may overlap other human CGI; FIREWACh N=8733, CapStarr-seq N=1149), d) 'conserved enhancer' elements overlapping a H3K4me1/H3K27Ac peak or enhancer HMM category in any human cell line) (FIREWACh N=48977, CapStarr-seq N=4758) or e) other mouse elements screened but without either enhancer feature in humans (FIREWACh N=35338, CapStarr-seq N=1202). We find that in both assays mouse CGI, regardless of their conservation status, exhibit greater activity than other enhancers. More than 20% of mouse CGI, whether they were conserved as CGI in humans or not, were recovered in FIREWACh assays, compared to ~10% of conserved enhancers or ~5% of mouse-specific enhancers (Supp. Fig. 2A). However, those that represent ECGI in humans are overrepresented relative to those representing classical enhancers (Conserved ECGI vs. Conserved Classical Enhancers OR=1.82, p= 0.023, Fishers Exact) and those that are conserved as CGI between human and mouse are overrepresented relative to those that have lost CGI characteristics (Conserved ECGI vs. Nonconserved ECGI OR= 2.71, p=8.16E-9, Fishers Exact). Similar findings were observed in CapStarr-Seq assays (Supp. Fig. 2B,C), wherein mouse CGI exhibited greater ability to drive transcription than CpG-poor enhancers, and those representing EGI were more active than classical enhancers (Conserved ECGI vs. Conserved Enhancers p=6.6E-4, Mann Whitney U), and were more likely than such elements to be defined as strong or weak, rather than inactive (strong or weak vs. inactive, Conserved ECGI vs. Conserved Enhancers OR=4.93, p= 0.0027, Fishers Exact; Conserved ECGI vs. Nonconserved ECGI OR= 4.36, p=0.0305, Fishers Exact).

Supplemental Figure Legends:

Supplementary Figure 1 A) Distribution of the fraction of analyzed cell lines (N=120) in which the loci in each class exhibit enhancer activity (overlap an annotated H3K4me1/H3K27Ac peak in that cell type). B) Stable or Unstable transcript pairs as defined in K562 cells (4) were intersected with each class of CGI (see Methods). Shown is the fraction of each set of regions overlapping stable, unstable or mixed transcript pairs. C) The normalized fraction of all such detected pairs in each class of CGI.

Supplementary Figure 2 A) Shown is the fraction of each set of elements (see Supplemental Results above) that were recovered (able to drive significant reporter GFP expression) in the FIREWACh assay. B) Shown is the distribution in the CapStarr-Seq enhancer activity score

(Fold Change [FC] in RNA expression in transfected cells vs. DNA copy number in input enhancer reporter plasmid libraries) for each set of genomic elements. C) The fraction of each set of genomic regions meeting the criteria in CapStarr-Seq to be called inactive (FC<1.5), weak (FC 1.5-3), or strong (FC >3).

Supplementary Figure 3: Distribution of the ChIP-seq density (reads/kb) for the indicated chromatin feature as measured in H1ESC cells (A) or K562 cells (B) among ECGI or classical enhancers active in the indicated cell type. Supplementary Figure 4 Overlap of ECGI active in MCF7 or HMEC cells, as defined by overlapping H3K4me1/H3K27Ac peaks and meeting an enhancer chromHMM definition (see Methods).

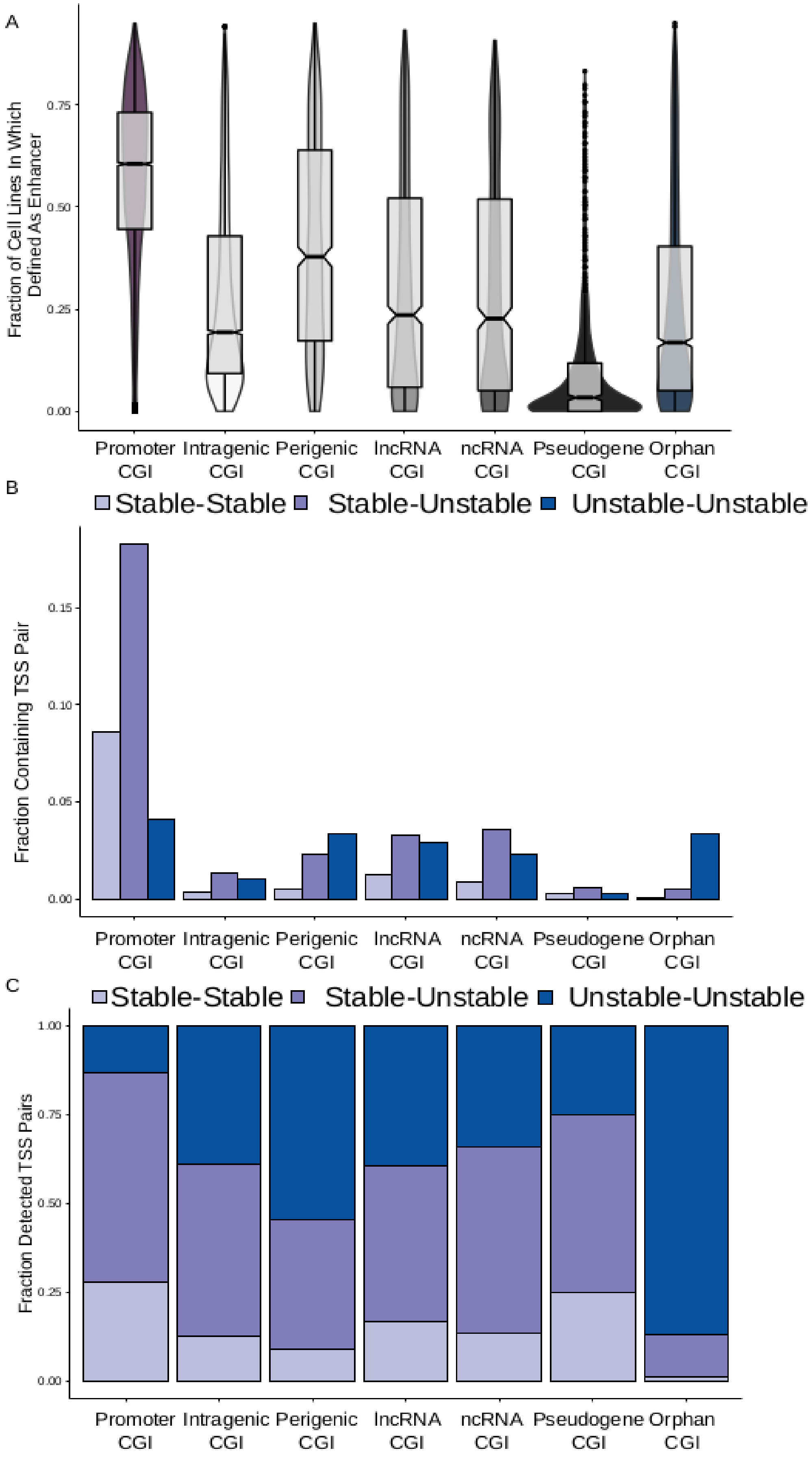
Supplemental References

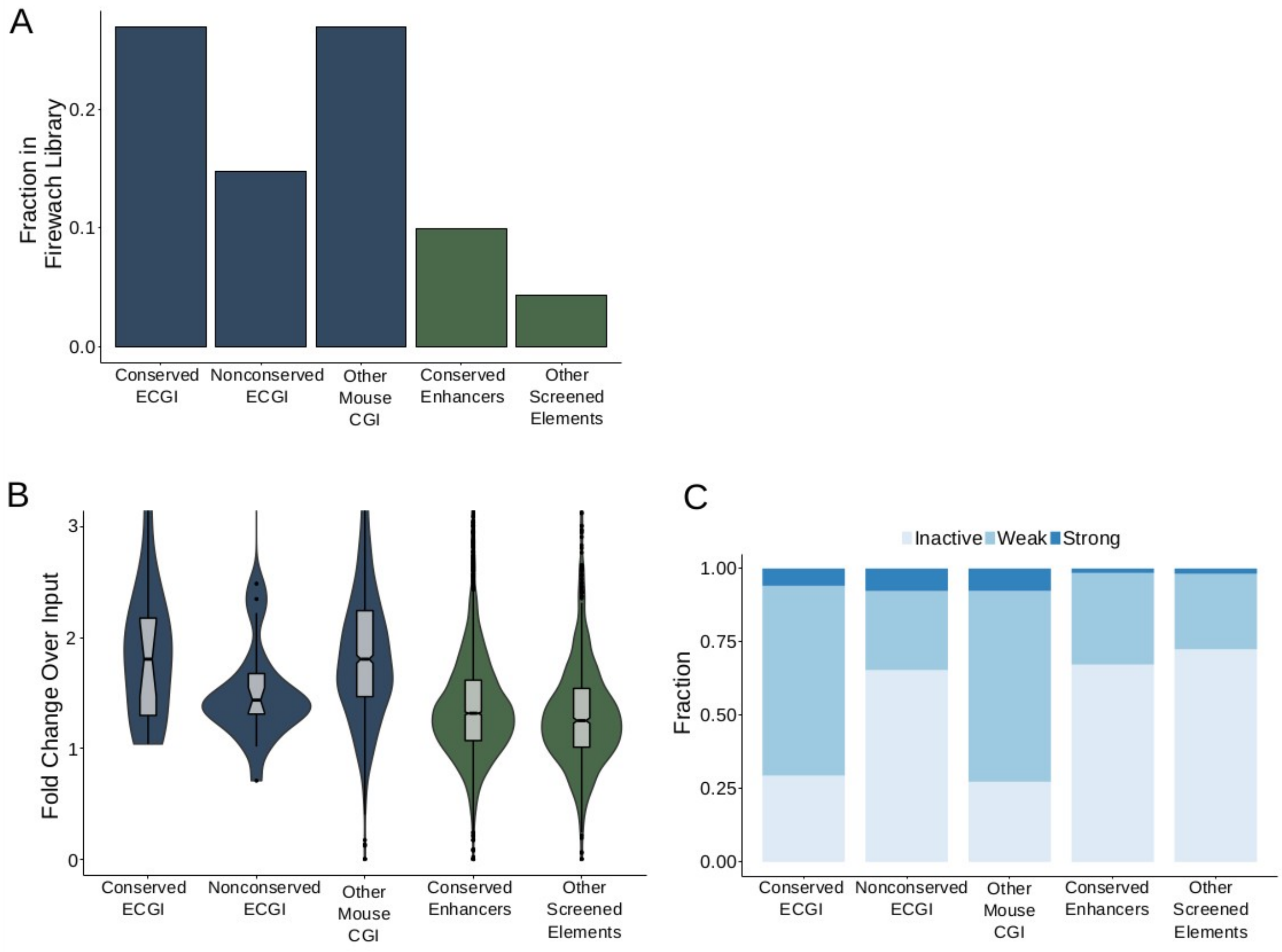
1. Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, et al. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nature methods. 2014;11(5):559-65.

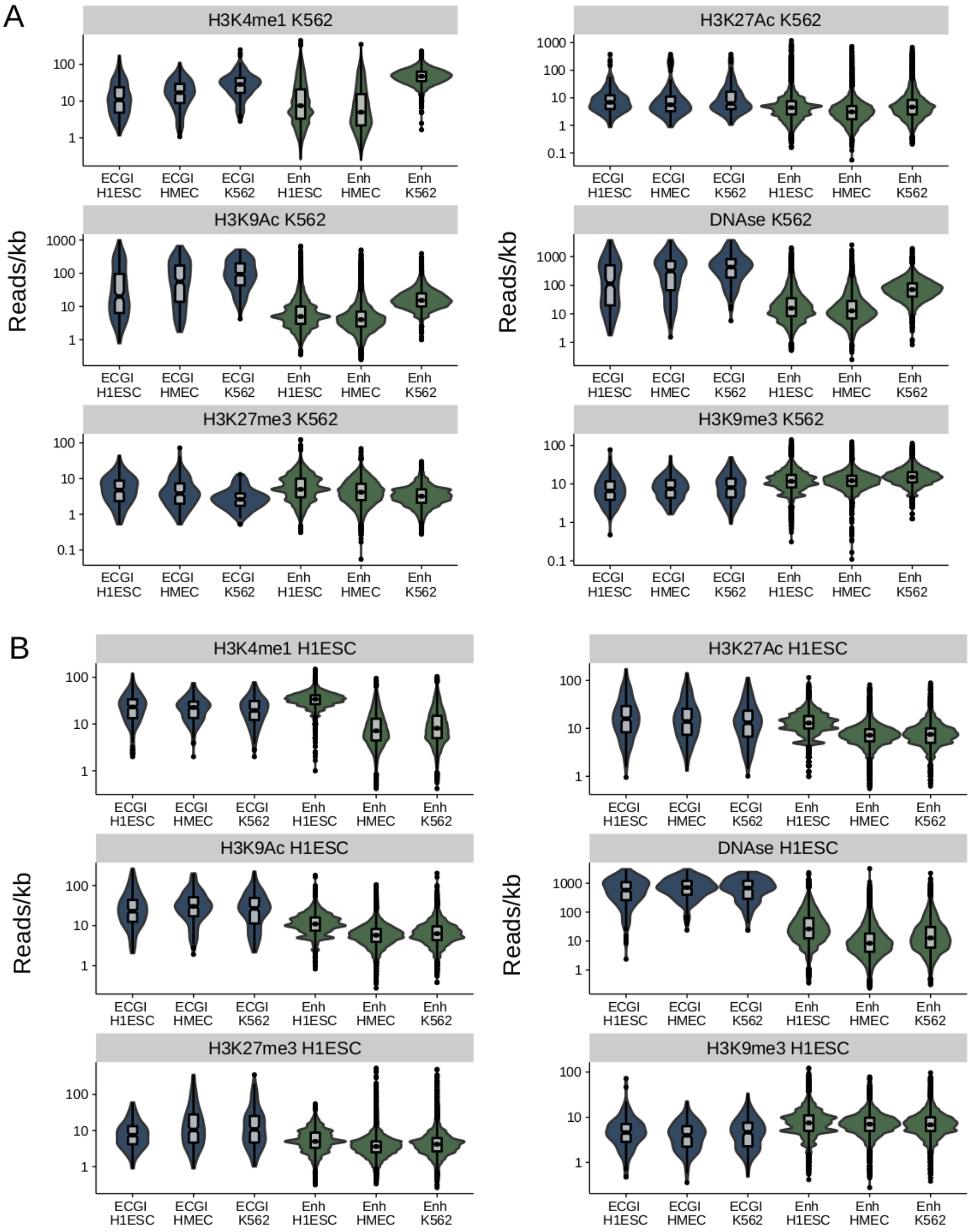
2. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339(6123):1074-7.

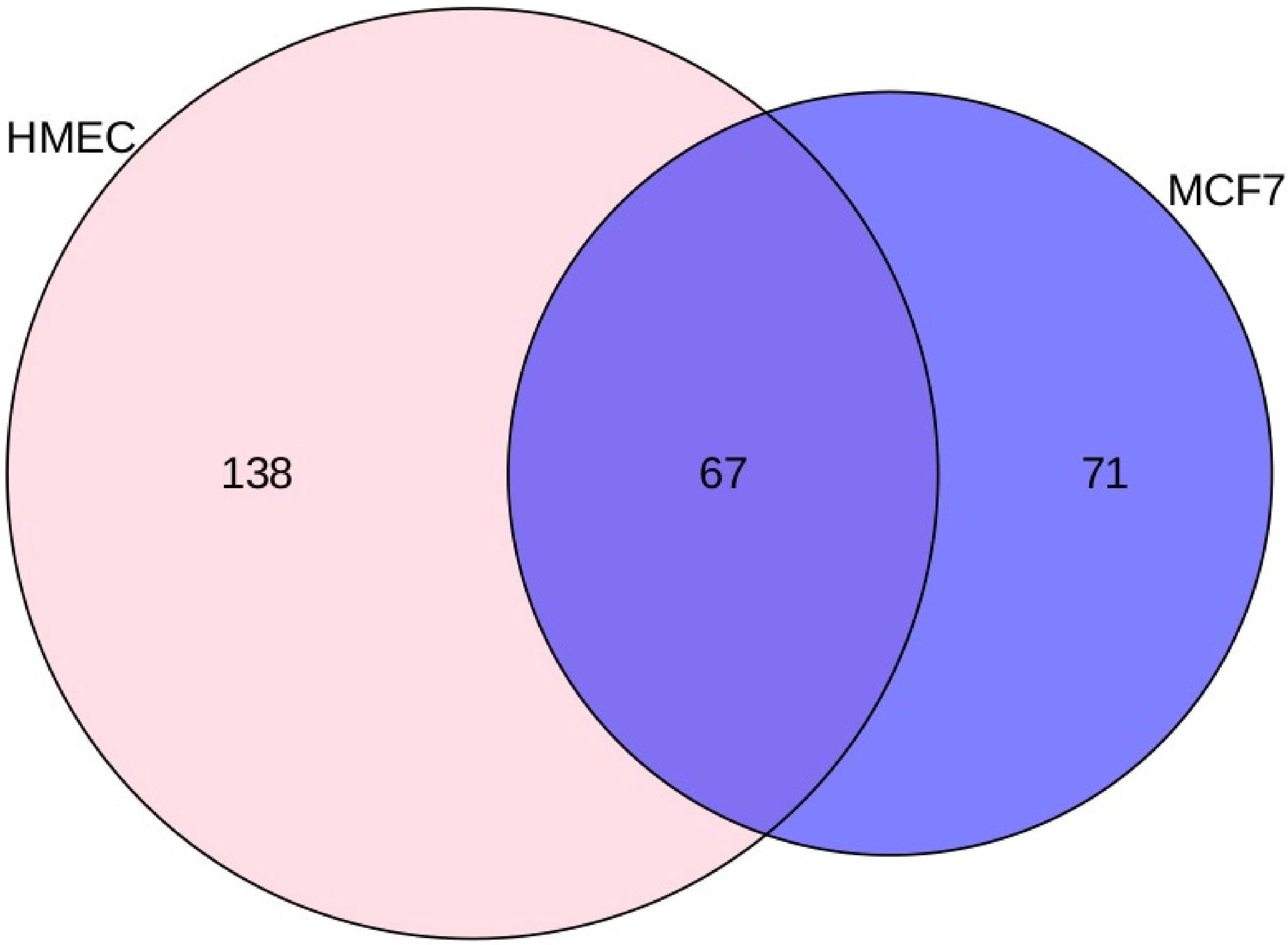
3. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LT, Fernandez N, et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nature communications. 2015;6.

4. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics. 2014;46(12):1311-20


A
Fraction of Cell Lines In Which Defined As Enhancer
0.00
0.25
0.50
0.75
Promoter CGI
Intragenic CGI
Perigenic CGI
lncRNA CGI
ncRNA CGI
Pseudogene CGI
Orphan CGI
B
Stable-Stable
Stable-Unstable
Unstable-Unstable
Fraction Containing TSS Pair
0.00
0.05
0.10
0.15
Promoter CGI
Intragenic CGI
Perigenic CGI
lncRNA CGI
ncRNA CGI
Pseudogene CGI
Orphan CGI
C
Stable-Stable
Stable-Unstable
Unstable-Unstable
Fraction Detected TSS Pairs
0.00
0.25
0.50
0.75
1.00
Promoter CGI
Intragenic CGI
Perigenic CGI
lncRNA CGI
ncRNA CGI
Pseudogene CGI
Orphan CGI


A
Fraction in Firewach Library
0.0
0.1
0.2
Conserved ECGI
Nonconserved ECGI
Other Mouse CGI
Conserved Enhancers
Other Screened Elements
B
Fold Change Over Input
0
1
2
3
Conserved ECGI
Nonconserved ECGI
Other Mouse CGI
Conserved Enhancers
Other Screened Elements
C
Inactive
Weak
Strong
Fraction
0.00
0.25
0.50
0.75
1.00
Conserved ECGI
Nonconserved ECGI
Other Mouse CGI
Conserved Enhancers
Other Screened Elements


A
H3K4me1 K562
H3K27Ac K562
H3K9Ac K562
DNAse K562
H3K27me3 K562
H3K9me3 K562
Reads/kb
0.1
1
10
100
1000
ECGI H1ESC
ECGI HMEC
ECGI K562
Enh H1ESC
Enh HMEC
Enh K562
B
H3K4me1 H1ESC
H3K27Ac H1ESC
H3K9Ac H1ESC
DNAse H1ESC
H3K27me3 H1ESC
H3K9me3 H1ESC


HMEC
MCF7
138
67
71

Chapter V: Therapeutic Inhibition of DNA methylation Resets the Cancer Transcriptome

Joshua SK Bell, Priya Kapoor, Paula M. Vertino

Contributions: JB and PK performed the experiments. JB analyzed the data and wrote the manuscript.

Abstract

DNA methylation is a key regulator of transcription in mammalian cells. Aberrant DNA methylation changes acquired during tumorigenesis are responsible for the silencing of tumor-suppressor genes, modulation of enhancer activity, and unleashing of endogenous retroviruses (ERVs). Epigenetic therapy inhibiting DNA methylation decreases the growth rate and invasive potential of solid tumor cells, and is a standard treatment for certain hematological cancers. We examine the impact of drug-induced demethylation on the nascent transcriptome of breast cancer cells using Precision Run-on Sequencing (ProSeq) to document that loss of DNA methylation results in pervasive transcriptional changes including the down-regulation of over-expressed oncogenes due to loss of gene body methylation, reactivation of genes silenced by promoter methylation, global enhancer activation, especially among enhancers hypermethylated in breast cancer, unmasking of cryptic intragenic promoters, and potent induction of telomeric repeats and certain ERV families. Critically, many of these transcripts remain highly paused or are unstable and thus have evaded detection by traditional RNA-Seq. These findings suggest new mechanisms by which epigenetic therapy reprograms cancer cells and asserts novel roles for DNA methylation in regulating transcription more broadly.

Introduction

DNA methylation is known to modulate the transcriptional program of both normal cells during lineage specification and in cancer cells during tumorigenesis and progression. Methylation of cytosines in a CpG dinucleotide context is a critical epigenetic mark with conserved roles across vertebrates and a key target of epigenetic therapy in cancer treatment [1]. Yet, the full role of DNA methylation in regulating transcription, and how inhibition of DNA methylation yields transcriptional changes that diminish cancer growth rates and invasive potential, remain incompletely elucidated.

DNA methylation is known to play important roles in preventing transcriptional initiation at promoters, both naturally in the silencing of lineage-restricted genes and pathogenically in the silencing of tumor-suppressors [2]. In mammals, a majority of promoters are marked by CpG Islands (CGI). CGI are distinct chromatin domains characterized by DNA hypomethylation, CpG density, high levels of bidirectional transcriptional initiation, and histone modifications associated with active expression: H3K4me3, H3K9/27Ac, for example. Most CGI in the genome remain unmethylated across tissues, a state which is permissive to transcriptional initiation, but gene expression level is further dependent on other factors, especially release from RNA Polymerase II (Pol II) pausing [3]. Following transcriptional initiation, Pol II pauses after elongating ~20-50 nucleotides and awaits release by recruitment of active PTEF-b [4]. At CGI promoters, Pol II also pauses distally at 3' CGI boundaries. We recently demonstrated that the degree and position of pausing correlates with GC-skew, the relative frequency of guanines to cytosines on the coding strand [3]. The relief of CGI promoter methylation by DNA methyltransferase inhibitors like deoxyazactydine (DAC) at key tumor-suppressors is well-documented, and thought to play a key role in their activity .

In contrast to the repressive role of DNA methylation in promoters, highly expressed genes actually tend to have methylated gene bodies, and in many organisms levels of DNA methylation in the gene body positively correlate with expression. Intragenic DNA methylation is thought to silence spurious promoters and repetitive elements or retroviruses present in genes. Exons also tend to be more methylated than introns, and DNA methylation may play a role in splicing regulation. Indeed, treatment with the DNMT inhibitor decitabine results in changes to exon inclusion in cells [5]. As well, many genes rely on DNA methylation for full expression: DAC treatment lowers the expression of many genes through gene-body demethylation that cannot regain their endogenous expression following drug withdrawal without DNMT3A [1].

Transcriptional initiation by Pol II is not limited to protein-coding mRNAs. At genes, mRNAs are typically paired with an upstream antisense transcripts, the roles of which are unclear but are thought to maintain an open transcriptional state at promoters. Similarly, Pol II initiates birectionally at enhancers, but both eRNA (enhancer RNA) transcripts are typically unstable. Enhancers are genomic regions that act at a distance to promote gene transcription independent of position or orientation. Enhancers are similar to promoters in that their levels of DNA methylation are inversely correlated with their activity and eRNA levels [6–9]. Indeed, we recently identified a subset of intergenic CGI that act as enhancers, and possess strength concomitant with intense hypomethylation rarely observed at CpG-poor enhancers. Enhancer methylation levels can change drastically during oncogenesis and during metastasis, a process known to effect key oncogenes like *KIT.* However, the extent to which enhancer activity is changed by DAC treatment has not been extensively studied[10].

DNA methylation also plays key roles in silencing constitutive heterochromatin, regions of the genome enriched in repeats, including centromeric and telomeric satellites, retrotransposons and endogenous retroviruses, that must be kept silent and compact in all cell types to protect genomic integrity from mutation and mitotic errors. Indeed, recently it was shown that the activation of certain endogenous retroviruses (ERVs) by AZA/DAC resulted in ERV dsRNA accumulation in the cytosol that triggers a signaling cascade mediated by the dsRNA sensor MAVS to dampen growth rate in cancer cells [11].

Several studies to date have utilized RNA-Seq to interrogate the changes to transcription induced by demethylation. However, RNA-Seq measures only processed transcripts of sufficient steady-state abundance and is blind to Pol II pausing dynamics, eRNAs, spurious transcripts, and often any non-polyadenylated transcript like ERVs. Thus, in order to determine the relationship between nascent transcription and DNA methylation, we utilized Precision Run-on Sequencing (ProSeq) [12]to map engaged Pol II genome-wide while treating with the DNA methylation inhibitor DAC.

Results

Decitabine treatment results in pervasive hypomethylation and transcriptional changes

To determine how decitabine alters transcription in cancer cells during therapy, and more broadly address how DNA methylation regulates nascent transcription, we treated MCF7 breast cancer cells with a single 300nM dose of decitabine. After three days, we assayed DNA methylation using the Illumina Infinium Epic Array [13], and assayed nascent transcription using ProSeq, both in biological duplicates. The Epic array assays over 850 thousand CpG sites with extensive coverage of transcripts and regulatory elements (Figure 1 A, B). In both replicates, we documented pervasive hypomethylation, with an average 30% decrease in Beta value, a proxy for percent methylation. Overall, ~650 thousand probes exhibited significant hypomethylation (FDR <.05), with ~450 thousand of these decreasing by at least 0.2 Beta. To determine how DNA methylation and transcription change in various epigenomic compartments we used chromHMM to partition the MCF7 epigenome into distinct chromatin states. We combined our mock-treated ProSeq data with ENCODE MCF7 ChIP-Seq data for twelve marks to divide the genome into Strong & Weak/Poised Promoters, Transcriptional Transition & Elongation, Weak Transcription, Strong & Weak Enhancers, Insulators, Polycomb-Silenced, and Heterochromatin, defined by their enrichment in histone modifications/variants and proximity to various genomic features (Supplementary Figure 1 A-C, see Methods). We first compared the distribution of compartments across the genome as a whole, versus CpGs in the genome. CpG sites are the primary sites for DNA methylation, and are depleted relative to other dinucleotides in the genome because of the mutagenicity of methylcytosine, which is prone to deaminating to thymine over evolutionary time. As such, CpGs are concentrated where there is a selection for them to remain unmethylated or where cytosine identity itself is required (for coding or TF binding, for example): particularly in promoters, gene bodies, and insulators (Supp. Fig 1D). Comparing the genomic CpG distribution to those sites assayed by the array reveals that the array covers each chromatin state well, but is targeted towards promoters and enhancers given their known functional roles, at the expense of heterochromatic sites.

We next compared how the methylation and transcriptional activity of CpGs in each chromatin state were altered by decitabine treatment (Figure 1C, D). As expected, Strong Promoters exhibit very low initial methylation and high levels of transcriptional, neither of which was changed by decitabine treatment. In contrast, Weak/Poised promoters are modestly methylated and less transcribed, but tend to be activated by demethylation. Transcribed regions, primarily gene bodies, are usually highly methylated and may both lose and gain tag density upon treatment, the mechanisms of which we investigate further below. Enhancers are also sites for transcriptional initiation of eRNAs. Enhancers tend to be moderately methylated, with activity correlated with the degree of hypomethylation. Consistent with this concept, Strong Enhancers displayed lower methylation and higher transcription prior to treatment than Weak Enhancers, but both enhancer classes exhibit significant gains in expression upon demethylation. Interestingly, insulators tend to be highly transcribed and moderately methylated, and decrease slightly in ProSeq tag density upon demethylation. Heterochromatin, compact and silenced by DNA methylation, loses substantial methylation upon treatment and undergoes broad transcriptional activation.

Pol II Initiation at Silent CpG Island Promoters is Reactivated by Demethylation

Promoter CGI are the most well-studied genomic elements in terms of DNA methylation. While generally unmethylated, promoter CGI can be hypermethylated during development or carcinogenesis, which represses the associated gene, often a tumor-suppressor. However, hypomethylation is merely permissive to gene expression: levels of transcriptional initiation, release from Pol II pausing, and elongation rate also play essential roles in determining total stable, full-length mRNA production. Promoter CGI tend to possess broad, bidirectional transcription, and, in addition to classic promoter-proximal pausing, we previously identified a distinct Pol II pausing point at the barriers of CGI, known as distal pausing [3]. We found that most CGI promoters tend to predominately exhibit either proximal or distal pausing, and we define these as Proximally or Distally paused genes, respectively. To precisely assess how transcriptional dynamics changed at promoter CGI, we examined those CGI with a single unique GenCode TSS sufficiently far from the downstream and upstream edges of the island to properly address both pausing sites and the relationship with upstream antisense transcription (at least 250bp upstream, and 500bp downstream). We determined the class of each promoter CGI by examining the tag densities in the proximal and distal pausing regions (TSS+250bp, and 3'CGI Edge -250, respectively), and in the gene body (3'CGI+250bp to TES). Genes in the lowest decile of expression in any region were considered to be Silent, and among active genes, we distinguished Proximal and Distal genes by which pausing region had higher tag density.

Among Proximal and Distal genes, we detect modest, thought consistent, gains in proximal tag density indicative of initiation increases, consistent with low DNA methylation in untreated cells (Fig 2A,B), and thus minimal DNA methylation loss. There were also concomitant increases in Pausing Index (dominant pause site tag density/ gene body tag density). However, this was largely linked to loss of gene body tag density (Figure 2C, see Figure 4). However, we observed massive increases in nascent transcription at methylated, Silent CGI, specifically in the proximal pausing region immediately downstream of the TSS. Examining tag densities more directly in Silent CGI, we observe that tag densities declines sharply at the point of promoter-proximal pausing and distal pausing, leading to relatively modest increases and low Pol II density in gene bodies. (Figure 2D). Thus, proximal and distal pausing both appear unaffected by demethylation, and serve to buffer increases in initiation.

DNA Hypomethylation Increases Transcriptional Initiation at Promoters, but Does Not Relieve Pol II Pausing

While the silencing of CGI promoters by DNA hypermethylation is well-established, less attention has been given to the role of methylation at genes lacking a CGI. We find that most such genes exhibit moderate to heavy promoter methylation, although a subset are unmethylated. Consistent with a role for methylation repelling Pol II initiation, these unmethylated genes display much greater expression than methylated genes, even as most of this transcription is also bound by Pol II pausing (Fig 3A,B). Like at active unmethylated CGI, we observed modest increases in initiation at this class of genes upon DAC-induced demethylation (Fig 3C). The degree to which demethylation impacted transcriptional initiation depended on the starting methylation levels surrounding the TSS. Much more significant initiation increases were observed at genes with high promoter methylation, with moderately methylated genes displaying moderate increases in initiation. In contrast to genes with CGI, which display much higher initiation levels more broadly, we observed increases in the gene body tag density, suggestive of higher full-length mRNA production at these genes. However, generally the pausing index of these genes increased (Fig 3D). These data represent further evidence that DNA methylation is a master negative regulator of transcriptional initiation at promoters, especially at CGI, but not does appear to regulate Pol II pausing.

Loss of Gene Body Methylation Preferentially Downregulates Highly Expressed Genes

DNA methylation has been assigned opposite roles at promoters versus at gene bodies; at promoters methylation impedes transcription, while at gene bodies methylation facilitates transcription [1]. Broadly, we are able to confirm these two principles. In order to directly address sense vs. antisense transcription, we limited this analysis to Gencode protein-coding genes that do not overlap any annotated antisense transcript (N=10,320). Transcribed genes do tend to exhibit much less promoter methylation than do silent genes, though exceptions abound (Fig 4A). Also, while highly transcribed are hypermethylated in their gene bodies, many or most lowly transcribed genes also exhibit substantial gene body methylation (Fig 4B). Thus, comparing promoter and gene body methylation across genes reveals that the majority of genes have unmethylated promoters and methylated gene bodies, regardless of expression level (Fig 4C). Parsing genes by promoter methylation reveals that those with methylated promoters tend to be upregulated by decitabine treatment, while those with unmethylated promoters tend to be downregulated, in spite of their increased Pol II initiation levels (Fig 4D). However, parsing genes by gene body methylation reveals that only genes with the highest methylation levels tend to be downregulated by decitabine treatment (Fig 4E). The fact that genes with CGI promoters tend to lose expression, while genes without CGI (which are expressed at much lower levels) are upregulated led us to examine how absolute transcriptional levels (gene body tag density) was linked to sensitivity to demethylation. Thus, we examined those genes with <25% promoter methylation and >75% gene body methylation (most genes), in order to control for loss of promoter methylation. We found that initial transcriptional levels appear to be the main driver of sensitivity to demethylation, with highly expressed genes being the most sensitive to down-regulation by DAC treatment (Fig 4F). These data suggest that gene body methylation plays a critical role in permitting active elongation of Pol II across gene bodies. Furthermore, while methylation is found at most genes, it is only necessary for full expression among the most highly transcribed.

Antisense Intragenic Transcription is Induced by Methylation Loss

Surprisingly, we also found that antisense transcription is abundant in genes lacking annotated antisense transcripts (Fig 4G). In general, antisense transcription is correlated with sense transcription. However, decitabine appears to selectively upregulate antisense transcription, while downregulating sense transcription, leading to overall decreases in sense/antisense ratios in most genes. Parsing genes by initial gene body methylation levels reveals a direct correlation between initial methylation and gains in antisense tag density (Fig 4H). Given that high levels of sense transcription are dependent on high DNA methylation levels, we also investigated if high antisense transcription levels are as well. Indeed, those genes with the most antisense transcription were also prone to antisense downregulation upon DAC treatment, further evidence of the dependence of efficient elongation on DNA methylation, regardless of strand.

To address the hypothesis that DNA methylation directly prevents initiation of cryptic antisense promoters in methylated gene bodies, we used Homer [14], which looks for regions of high tag density (putative TSS) next to regions of lower tag density (putative transcript bodies) to annotate nascent transcripts. Surprisingly even in untreated MCF7, most (57%) genes without an annotated antisense trancript were found to have one by ProSeq. These may represent highly unstable transcripts not detectable by RNA-Seq. Upon decitabine treatment, we were able to detect novel antisense transcripts gained at 17% of genes, while in contrast, only 6.5% of genes lost an antisense transcripts (Fig 4J). In contrast, 72% of genes had a detectable sense transcript in untreated cells, and a similar number gained and lost such transcripts (11.8% vs 10.19%, respectively). Examining the promoter and gene body methylation of genes that gained or lost transcripts, we find that as expected, genes that gained sense transcripts tended to have promoter methylation, whereas those lost sense transcripts tended to have unmethylated promoters. However those genes sensitive to gain or loss of antisense transcripts had unmethylated promoters, suggesting a dependence on gene body methylation instead (Fig 4K).

The reliance of highly expressed genes on DNA methylation suggests that oncogenes highly over-expressed in breast cancer may be particularly sensitive to down-regulation by

DAC treatment, while those silenced by promoter methylation may be upregulated. To investigate this, we used the TCGA database of RNA-Seq from normal breast tissue (N=84) and tumor samples (N=784) and defined genes that were at least 2-fold up or down-regulated in the tumor samples with a strict FDR of <.001 (N=1721 over-expressed, N=1347 under-expressed). Indeed, we found that genes over-expressed in breast cancer were likely to be down-regulated by DAC treatment, while those under-expressed tended to be up-regulated by DAC treatment (Fig 4L,M). Together these data suggest that the transcriptome is returned to a more normal state by DAC-induced hypomethylation.

Decitabine Results in Pervasive eRNA Induction

Given the association between enhancer hypomethylation and activity levels, we next addressed how DAC treatment affected eRNA production. To define active enhancers, we used the common definition of overlapping H3K4me1 and H2K27Ac peaks (ENCODE[15]). To generate a database of putative enhancers active in MCF7 or in other cell lines, we examined these marks in over one hundred cells lines in ChIP-Seqs performed by the ENCODE and Roadmap Epigenomics projects. We further divided enhancers into classical enhancers and enhancer CGI (ECGI) based on overlap with UCSC annotated CGI. We recently demonstrated the ECGI compose a distinct class of intergenic enhancers with highly elevated CpG density, GC content, minimal DNA methylation, euchromatic features including higher H3K4 methylation states and histone acetylation, broader expression across tissues, more genomic contacts, and more potent ability to drive a reporter gene than classical enhancers (In press, Chapter IV).

Consistent with previous results, we found that active classical enhancers exhibited lower DNA methylation than inactive classical enhancers (Fig 5A). This pattern holds true for

ECGI as well, which when active display even less DNA methylation than active classical enhancers, as well as much higher eRNA production (Fig 5B). Broadly, we found transcriptional upregulation of most enhancers correlated with their initial DNA methylation levels (Fig 5D). Enhancers with less than 25% average DNA methylation were not likely to be upregulated, but those with moderate or high methylation tended to gain transcription (Fig 5C). ECGI in particular were prone to upregulation, consistent with a greater dependence on CpG hypomethylation in permitting transcription of these elements.

Epigenetic therapy is thought to selectively reset the transcriptome to a more normal state. Thus, we next examined whether enhancers active in HMEC, but that had lost either H3K4me1 or H3K27Ac in MCF7 cells (apparent cancer-related decommissioning), were more likely to be activated than other inactive enhancers (those apparently silenced during normal lineage-specification). Indeed, we found that HMEC enhancers and ECGI were both much more prone to upregulation than other inactive enhancers, especially if they had acquired high methylation levels in MCF7 cells (Fig 5E). This suggests that hypermethylation-mediated silencing of enhancers in cancer is reversible to some extent by DAC therapy, and that enhancers from the normal cell type of origin may be primed for activation.

Upstream Antisense Transcription is Induced by Demethylation

Paired transcriptional initiation appears to be a near-universal feature of promoters. While mRNAs are by definition stably transcribed, they are typically paired with upstream antisense transcripts (uaRNA), which like eRNA, are highly unstable and not capture by RNA-Seq. Given the increases we observed in mRNA initiation levels, we examined uaRNA at both protein-coding genes and lncRNAs. First, we observed that uaRNA (tag density in the 250bp upstream of annotated TSS) and mRNA initiation levels (tag density in the 250bp downstream of annotated TSS) are tightly correlated, consistent with previous studies[16] (Supp Figure 2A). Just as CGI promoters have higher mRNA levels than other promoters, they also possess higher uaRNA levels (Supp Fig 2B). Consistent with a similar role for DNA methylation in regulating sense and antisense initiation, changes in uaRNA and paused mRNA levels were well correlated upon demethylation, with most genes gaining Pol II initiation at both TSSs (Supp Fig 2C).

Decitabine Unleashes Transcription of Repetitive Elements and ERVs

Given the known roles for DNA methylation in suppressing repetitive elements, we next examined various classes of repeats, using the UCSC repeat database [17]. Among those elements with probe coverage on the Epic array, we note that most repeat families are mostly, but not fully methylated which was expected given that cancer cell lines in particular tend to display hypomethylation of repeats relative to normal primary cells (Fig 6A). However, most repeat families were roughly as likely to increase as to decrease in expression, with the exception of telomeric repeats, SVA repeats, and ERV1 (Fig 6B). We focused further on subfamilies of these repeats, finding that while both telomeric subtypes (TAR1 and REP522) were prone to upregulation, different members of the SVA (Fig 6C) and ERV1 (Fig 6E) families exhibited distinct sensitivities to activation by DAC. The ERV1 family is of particular interest because many members, including the most highly induced LTR12C (Fig 6F), is bidirectionally transcribed and can form the dsRNA known to trigger cytosolic receptors to induce MAVS-mediated reductions in cancer cell proliferation and invasion [18]. However, many other members of the ERV1 family exhibited gains and losses, or were largely insensitive to change. It in unclear why some elements should be more prone to upregulation than others, given that most are DNA methylated (and undergo similar levels of demethylatoin), but this may reflect differential redundancy of other heterochromatic silencing mechanisms like H3K9me3 or H4K20me3, for example.

Discussion

In this study, we have found that DNA methylation is a critical barrier to transcriptional initiation across the genome, while also necessary for productive Pol II elongation across highly expressed genes. Not only do we observe clear correlation between methylation and Pol II initiation levels, we demonstrate that loss of DNA methylation results in pervasive increases in initiation, not only at methylated CGI promoters, but at non-CGI promoters, enhancers, pseudogenes, lncRNAs, gene bodies, enhancers, and repetitive elements. Additionally, we find that these changes tend to preferentially reactivate genes and enhancers silenced by DNA methylation during cancer progression, as well as down-regulating genes aberrantly over-expressed in cancer, strong evidence for the model that epigenetic therapy serves to reset the cancer epigenome.

The reactivation of silent, methylated CGI has been a focus of DAC research for decades [19,20]. Indeed, we find that genome-wide such CGI are prone to robust increases in initiation, at least as gauged by Pol II activity in the promoter. Notably, though, even with large increases these newly unmethylated CGI do not gain initiation levels that even approach those of already active CGI, however they do gain much more initiation than methylated non-CGI promoters. This suggests that while indeed methylation is an important silencing mechanism at these CGI, relief of methylation is insufficient for full gene activation. These genes may require additional histone modifications or TF binding for full activity, an idea consistent with the synergistic effects of HDAC inhibitors and decitabine that may allow increased histone acetylation at DAC targets [21] .

Notably, this trend also extends to ECGI. While active ECGI are typically unmethylated, the subset with moderate methylation undergo large fold changes in transcriptional intensity (eRNA levels) upon DAC-induced demethylation. ECGI that are silenced by methylation undergo even greater increases, indeed the most of any set of genomic regions we analyzed directly. While classical enhancers were also prone to activation upon demethylation, their eRNA levels never approach that of CGI. Because active classical enhancers typically exhibit much more methylation than active (or even many inactive) ECGI, these data suggest that ECGI are much more reliant on a hypomethylated state than are classical enhancers. This could be due to the fact that ECGI have a much greater CpG density, and thus provide many more platforms for euchromatic chromatin modifiers (especially those CXXC domains or those that can form complexes with such proteins) [22]. Alternatively, because the transcription factors most enriched in ECGI tend to contain CpG sites in their binding motifs (Chapter IV), methylation may play a more direct role in preventing TF binding at ECGI vs. classical enhancers. Either or both of these mechanisms may also contribute to the susceptibility of silent CGI promoters to DAC-induced activation.

DNA methylation has also been linked to classical enhancer activity, with active enhancers typically displaying lower methylation levels than inactive enhancers [8,23], but no studies thus far have assayed eRNA production genome-wide while altering DNA methylation levels. With ProSeq, we were able to document massive induction of eRNA across enhancers already active in MCF7, as well as hundreds of thousands of others that would ordinarily be restricted to other lineages. Classical enhancers, even when active, still tend to exhibit significant methylation levels which do not approach those seen at CpG-poor promoters with similar methylation levels, possibly indicating that they have less sequence-based transcriptional potential. Nonetheless, methylation appears to play a critical role not only in modulating active enhancers, but silencing lineage-restricted enhancers. It may be necessary to buffer the initiation potential at active enhancers because they may titrate Pol II away from the promoters with which they form intimate contacts. Alternatively, too much eRNA production may exacerbate release from Pol II pausing and enhancer methylation may serve to preserve pausing regulation, at both the proximal and distal pausing sites (given that both appear to rely on enhancer contacts for release). Tight restriction of enhancer activity may also focus the enhancer on its preferred target gene while preventing promiscuous interaction with, and possible activation of, spurious targets that may be closer on the linear genome or in three-dimensional chromatin topology. The need to silence lineage-restricted enhancers with DNA methylation would appear to arise from a need to prevent activation of lineage-restricted genes. However, enhancer methylation may also serve to prevent the titration of methylation-sensitive transcription factors, transcriptional machinery including Pol II, and chromatin-modifying enzymes by elements with transcriptional potential.

DNA methylation also serves as a constant barrier to the transcription of various families of repetitive DNA, including endogenous retroviruses, pseudogenes, and countless other unannotated features with transcriptional potential [24,25]. Indeed, DAC appears to result in the induction of many families of repeats, especially telomeric heterochromatin, SVA repeats, and ERV1 elements. Cancer cells are already more prone than normal cells to genome instability and the mitotic errors that transcription in these regions contributes to, indicating they may be further destabilized by DAC-induced loss of repeat regulation. Regarding various retroviral families, we found that certain classes were far more prone to reactivation upon DAC-induced demethylation than others. Given the role of these elements in forming dsRNAs that trigger MAVS-mediated loss of proliferation rate and invasive potential, our work may help focus efforts targeted on understanding this response.

Surprisingly, we found that antisense transcription within gene bodies is common, and indeed correlates well with levels of sense transcription. Antisense transcription in gene bodies increased in most gene bodies with high initial levels, suggesting that genes often contain spurious transcription start sites that are silenced by DNA methylation. The consequences of spurious intragenic transcripts are unclear. The positive correlation between sense and antisense transcription, as well as the thousands of genes with annotated endogenously transcribed ncRNA (or even other genes), suggests that antisense transcription per se does not negatively impact mRNA elongation. We also failed to observe a correlation between increases in antisense transcription and loss of sense transcription (discussed below) upon DAC treatment, but the pervasive nature of these unstable ncRNA raises interesting questions. It is possible that gene body antisense RNA may perform similar roles as eRNAs in releasing Pol II pausing, or may be translated and interfere with cellular processes. Alternatively, it is tempting to speculate that transcription from the coding strand (antisense transcription) may help prevent it from annealing with the template strand of the mRNA and thus assist in elongation.

It has been widely observed that DNA methylation in gene bodies correlates positively with mRNA levels. Indeed, we find in ProSeq that there is a modest correlation between gene body Pol II activity and DNA methylation, but the majority of genes exhibit substantial DNA methylation regardless of their transcriptional level and methylation alone is poorly predictive of mRNA output. In general, the role ascribed to gene body methylation is the silencing of cryptic promoters [26], for which we do find abundant evidence. However, our analysis also revealed that most highly expressed genes lost gene body tag density upon hypomethylation, and indeed that there was a direct positive correlation between initial expression level and sensitivity to downregulation, especially among genes with highly methylated bodies (and unmethylated promoters to control for initiation effects). This suggests a model in which DNA methylation is actively recruited to genes broadly as they are transcribed (perhaps by H3K36me3), even at modest levels. The finding that gene bodies are particularly sensitive to remethylation following DAC-induced demethylation is further evidence of this [27]. At gene bodies, this DNA methylation does apparently serve to prevent cryptic initiation in genes, and active genes may be more prone to such events by virtue of being near or within active chromatin domains. It was also recently demonstrated in yeast, which lack DNA methylation, that genes have pervasive antisense RNAs embedded in gene bodies that is silenced by H3K36me3 [28]. This may also be true in humans, but DNA methylation apparently serves this function as well. However, methylation also is apparently required for high levels of expression. The mechanisms of this remain unclear, but methylation has been suggested to slow Pol II elongation, especially in exons, to allow sufficient time for splicing machinery to act [5]. Thus, it is possible that without methylation to impeded its progression Pol II proceeds too quickly and is forced to abort transcription. Alternatively, cryptic promoters originating on the same strand as the mRNA may manifest unstably bound Pol II complexes that may perturb the elongation of transcripts originating from the upstream promoter. Further experiments, particularly ones designed to capture Pol II elongation rates following demethylation, will be required to investigate these possibilities.

Epigenetics is one of the most promising new avenues for the development of novel cancer therapies, and DNA methylation inhibitors in particular have been vital as single agents, but also in combination with other drugs. In principal, epigenetic therapy is thought to work by reprogramming the cancer epigenome to a more normal state. In this work we have identified several mechanisms that support this model. We showed that the most highly expressed genes, especially those over-expressed in cancer, are the most sensitive to down-regulation by gene body hypomethylation. Conversely, ProSeq also revealed the activation of many genes silenced by promoter methylation in breast tumors. While this has previously been appreciated at many individual loci, RNA-Seq fails to unearth the larger increases in initiation, buffered by proximal and distal pausing, that DAC-induced demethylation yields. Unexpectedly, we also observed that lineage-specific enhancers silenced during cancer progression were also especially prone to reactivation. This apparent priming of HMEC enhancers may be due to retained chromatin modifiers at such loci, or a permissive nuclear environment possibly including TFs like ESR1 or FOXA1 that are found in both normal breast tissue and MCF7 breast cancer cells. Regardless of the mechanism, this activation of normal breast epithelial enhancers is likely important for re-expression or upregulation of genes silenced by enhancer activity loss in cancer. Indeed, it has only been recently appreciated that many of the changes in gene expression associated with oncogenesis or cancer progression are linked to enhancer DNA methylation changes [8], and this research is the first to reveal that the reversal of these changes is more specific to enhancers hypermethylated during oncogenesis, rather than those naturally silenced during lineage specification. We and others have previously shown that DAC-induced reversal of aberrant DNA methylation is more stable at normally unmethylated targets, and that cancer-related hypermethylation is unlikely to return [27]. Together, these data suggest that not only does decitabine specifically target aberrant methylation, but it does so stably, with critical implications for cancer treatment and the continued exploration of DNA hypomethylating agents as combination therapies.

Methods

Cell Culture and DAC Treatment

MCF7 cells were obtained from ATCC and cultured in DMEM supplemented with 10% FBS. Cells were plated in 25cm dishes at density of 3E5 cells per plate and were treated the next day with a single dose of 300nM DAC (Fisher) or mock-treated with an equal concentration of solvent (50% Acetic acid) in fresh media. Cells were harvested after 3 days of treatment. DNA and RNA were prepared using Qiagen AllPrep extraction kit. Two independent biological replicates were performed.

ProSeq

ProSeq was performed as previously described. Briefly, nuclei were isolated from treated cells [12]. Native nucleotides are washed away, while Pol II remains bound to chromatin. Transcription is then allow to continue in the presence of biotin-labeled NTPs allowing for single-based extension of nascent transcripts. RNA is then hydrolyzed with NaOH, and labeled species are captured using streptavidin beads. The 3' adapter is then ligated, followed by removal of the 5' cap with RNA 5′ pyrophosphohydrolase (RppH), and the 5' end repaired with T4 polynucleotide kinase (PNK). The 5' adapter is then ligated, followed by reverse transcription, and PCR amplification for library preparation. We performed single-end 50bp sequencing (HudsonAlpha) and obtained ~100-150 million reads per data set.

For data processing, adapters were trimmed using the Fastx toolkit [29] . Reads under 16bp were removed, and then remaining reads were mapped to hg19 using Bowtie2 [30]. Duplicate and non-uniquely mapped reads were removed using Picard [31]. Tag densities in genome regions were determined using the GenomicRanges R package[32]. Fold changes in genomic regions were calculated using DESeq2 [33]with default settings, except for library size.

To normalize for enrichment differences, library size values were normalized to the fraction of reads in each data set mapping to unmethylated CGI containing an annotated GenCode TSS where tag density is not expected to change significantly and this fraction did not appear to be impacted by treatment: Mock Replicate 1 9.1%, Mock 2 10.1%, DAC 1 10.7%, DAC2 9.9% (p=0.39 between treatments).

DNA Methylation

Genomic DNA was bisulfite converted and hybridized to the Illumina Infinium EPIC BeadChip by the Emory Integrated Genomics Shared Resource. The ChaMP Bionconductor package [34], with default settings and SWAN normalization, was used to produce beta values for each probe, and filter low-quality and SNP-containing probes. Density plots of methylation levels were made using ggplot2 [35], and heatmaps using the heatmap3 [36] package. Average methylation values in genomic features were calculated using the GenomicRanges R package.

ChromHMM

The chromHMM software was run using default standards with ENCODE ChIP-Seq datasets (Table 1) for 13 states which were annotated based on similarity to states as described in Ernst & Kellis [37]. Two states resembling Transcriptional Elongation were merged, as were a Heterochromatin state (lacking any analyzed marks) and one marked only by H3K9me3. The Repetitive/CNV class (showing enrichment for all marks) was omitted from analysis given the difficulty in uniquely mapping reads to these regions (see Supplementary Figure 1).

Figure 1: Pervasive Hypomethylation and Transcriptional Changes Across Genomic Compartments

A) Density plot and B) heatmap of DNA methylation levels (Beta values) in two replicates of mock and DAC-treated cells. C) Distribution of percent methylation (average of both replicates) in each chromHMM category in mock (green) and DAC-treated (blue) cells. ChromHMM categories are defined in 1kb windows based on 12 chromatin marks in MCF7 cells (with merging of identical adjacent windows). Boxes represent the 1rst and 3rd quartiles, with hinges at the median, and hinge width the 95% confidence interval. Lines represent the further data point within 1.5x the interquartile range. D) Distribution of log2 fold changes in ProSeq tag density (average of both replicates) in each chromHMM category between mock and DAC-treated cells.

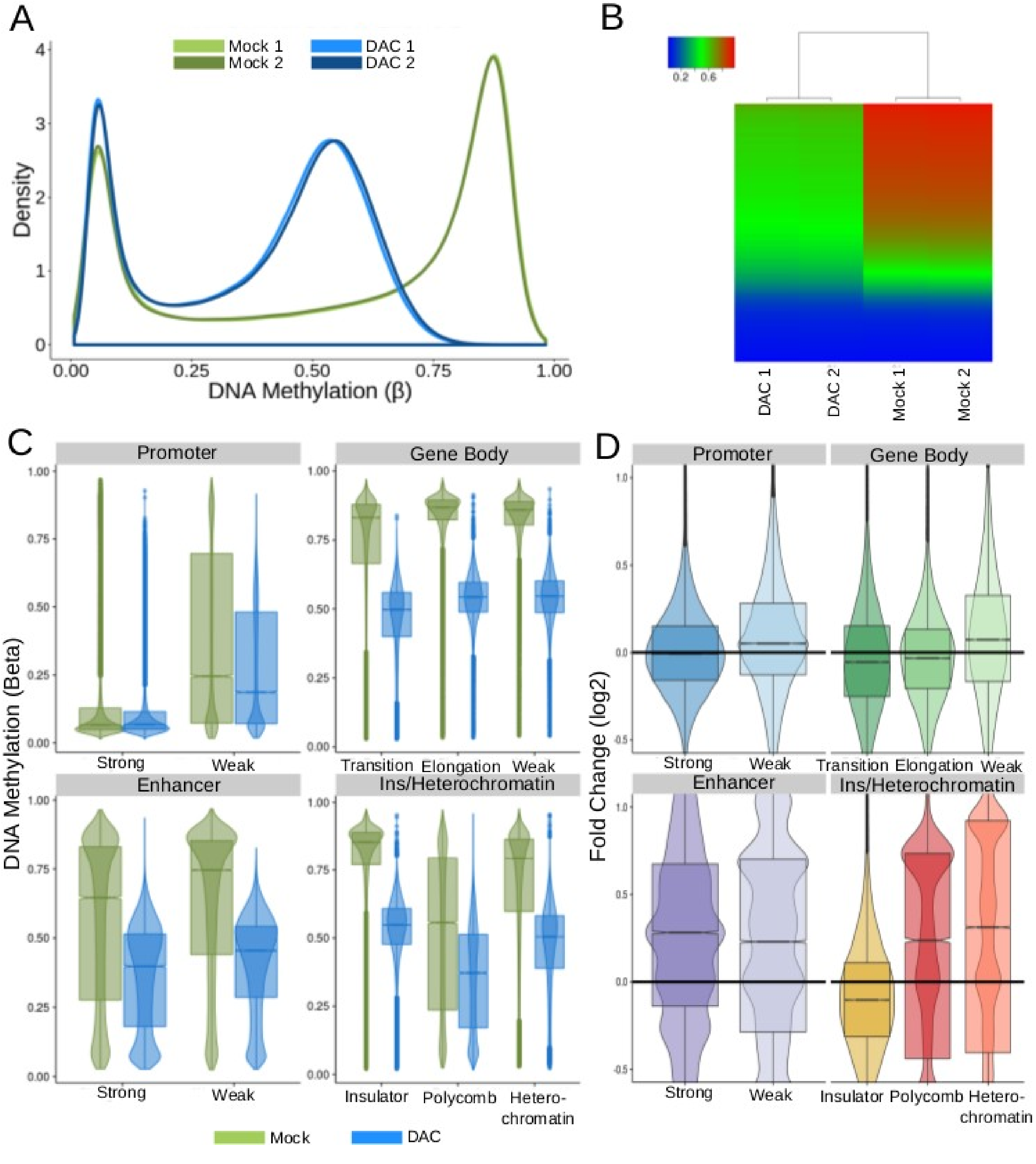

A
Mock 1
Mock 2
DAC 1
DAC 2
Density
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
DNA Methylation (β)
B
0.2
0.6
DAC 1
DAC 2
Mock 1
Mock 2
C
DNA Methylation (Beta)
Promoter
Strong
Weak
Gene Body
Transition
Elongation
Weak
Enhancer
Strong
Weak
Ins/Heterochromatin
Insulator
Polycomb
Hetero-chromatin
0.00
0.25
0.50
0.75
1.00
Mock
DAC
D
Fold Change (log2)
Promoter
Strong
Weak
Gene Body
Transition
Elongation
Weak
Enhancer
Strong
Weak
Ins/Heterochromatin
Insulator
Polycomb
Hetero-chromatin
-0.5
0.0
0.5
1.0

Figure 2: Pol II Initiation at Silent CpG Island Promoters Reactivated by Demethylation

A) Percent methylation across Proximal, Distal, and Silent CGI promoters. UCSC CGI with unique Gencode TSSs were parsed by average proximal tag density (TSS+250bp) and distal tag density (3'CGI boundary -250bp), with Proximal genes have more proximal tags, Distal genes with more distal tags, or Silent genes which were in the lowest decile of expression in either region, or in the gene body (3'CGI boundary +500bp to TES). CGI were independently scaled from the 5' CGI edge to the TSS, and from the TSS to the 3' CGI edge, and plotted with the flanking 2.5kb, oriented in the direction of mRNA transcription. B) ProSeq tag density for features as defined in A, with sense tags (on the same strand as the mRNA) plotted as positive values, and antisense tags plotted as negative values. C) Fold changes (log2) in ProSeq tag density for each region and set of genes as defined in A between mock and DAC-treated cells. D) Distribution of average ProSeq tag density in mock (green) or DAC-treated (blue) cells for each region of Silent genes.

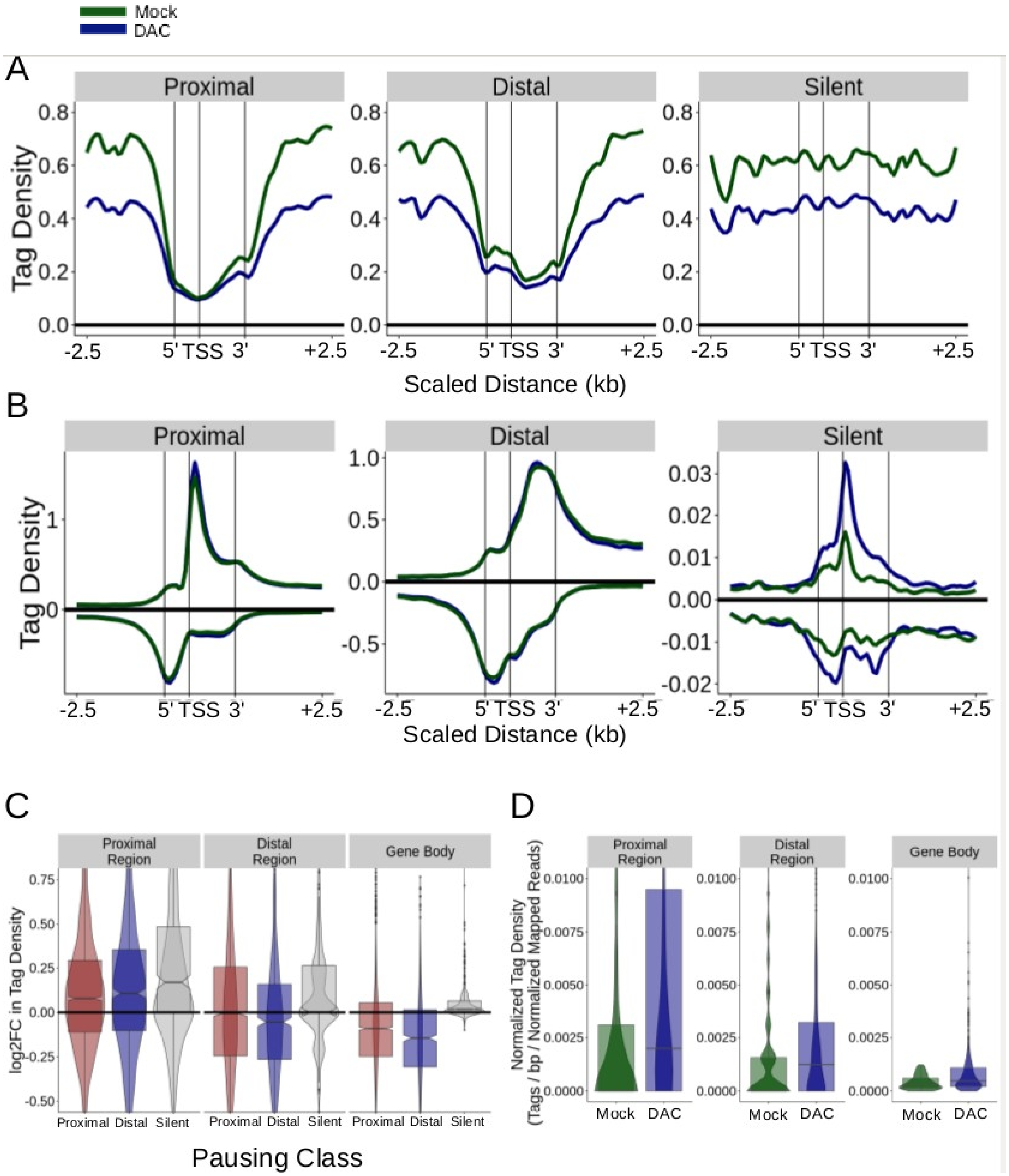

Mock
DAC
A
Proximal
Distal
Silent
Tag Density
0.8
0.6
0.4
0.2
0.0
-2.5
5'
TSS
3'
+2.5
Scaled Distance (kb)
B
Proximal
Distal
Silent
Tag Density
1
0
1.0
0.5
0.0
-0.5
0.03
0.02
0.01
0.00
-0.01
-0.02
-2.5
5'
TSS
3'
+2.5
Scaled Distance (kb)
C
Proximal Region
Distal Region
Gene Body
log2FC in Tag Density
0.75
0.50
0.25
0.00
-0.25
-0.50
Proximal
Distal
Silent
Pausing Class
D
Proximal Region
Distal Region
Gene Body
Normalized Tag Density
(Tags / bp / Normalized Mapped Reads)
0.0100
0.0075
0.0050
0.0025
0.0000
Mock
DAC

Figure 3: DNA Hypomethylation Increases Transcriptional Initiation at Promoters, but Does Not Relieve Pol II Pausing

A) Percent methylation or B) sense ProSeq tag density across Gencode protein-coding promoters without a CGI within 2kb. Genes were separated by average methylation in the 1kb region surrounding the TSS: Low <25%, Moderate 25-75%, High >75%, oriented in the direction of mRNA transcription. C) Distribution of average ProSeq tag density in the proximal region (TSS+250bp) or gene body (TSS+500bp to TES) in mock (green) or DAC-treated (blue) cells, as defined in A. D) Distribution of pausing index (PI, proximal tag density divided by gene body tag density) among genes as defined in A.

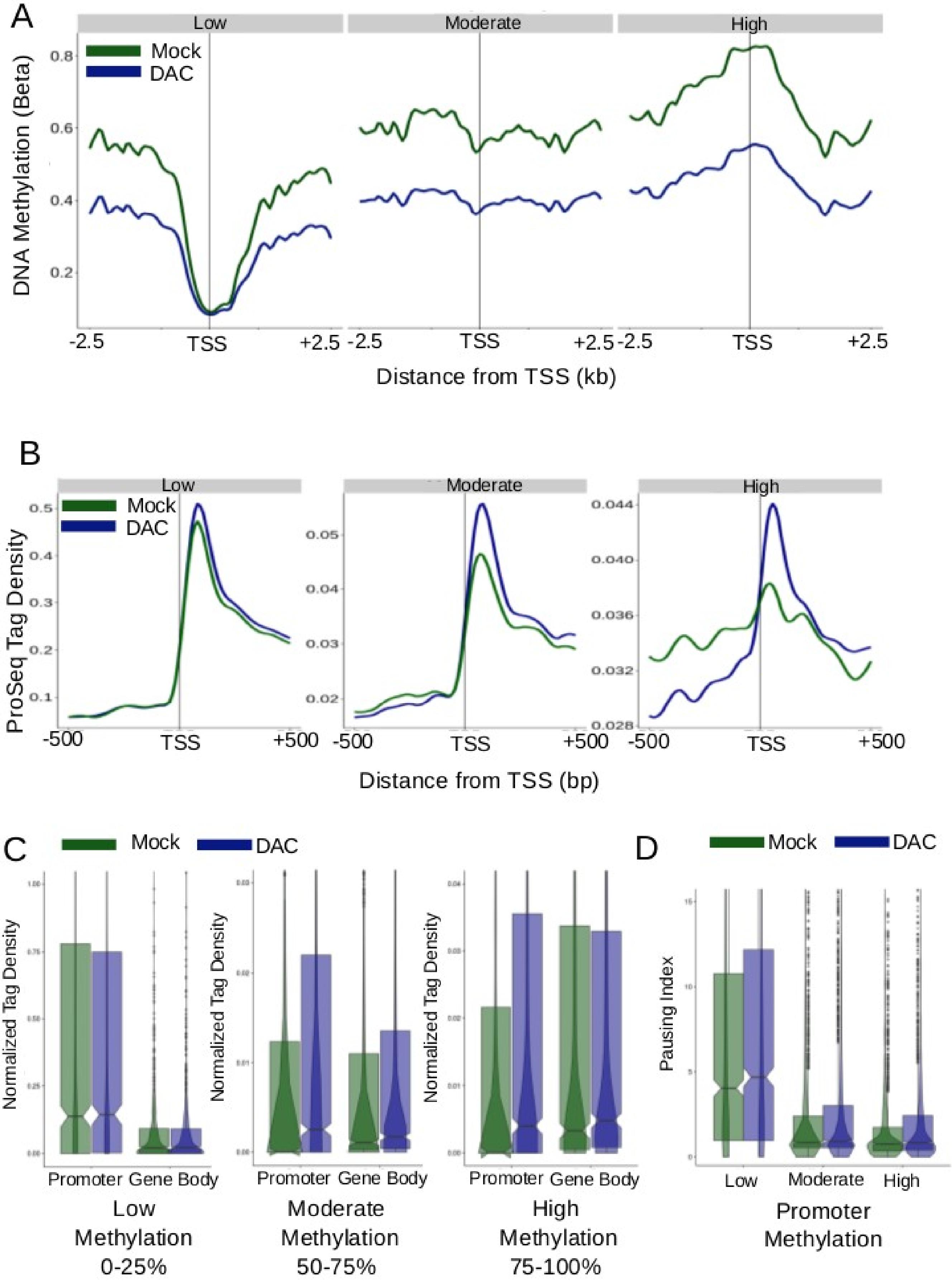

A
Low
Moderate
High
Mock
DAC
DNA Methylation (Beta)
0.8
0.6
0.4
0.2
-2.5
TSS
+2.5
-2.5
TSS
+2.5
-2.5
TSS
+2.5
Distance from TSS (kb)
B
Low
Moderate
High
Mock
DAC
ProSeq Tag Density
0.5
0.4
0.3
0.2
0.1
0.05
0.04
0.03
0.02
0.044
0.040
0.036
0.032
0.028
-500
TSS
+500
-500
TSS
+500
-500
TSS
+500
Distance from TSS (bp)
C
Mock
DAC
Normalized Tag Density
1.00
0.75
0.50
0.25
0.00
Normalized Tag Density
0.03
0.02
0.01
0.00
Normalized Tag Density
0.04
0.03
0.02
0.01
0.00
Promoter Gene Body
Low
Methylation
0-25%
Promoter Gene Body
Moderate
Methylation
50-75%
Promoter Gene Body
High
Methylation
75-100%
D
Mock
DAC
Pausing Index
15
10
5
0
Low
Moderate
High
Promoter
Methylation

Figure 4: Loss of Gene Body Methylation Preferentially Impacts Highly Expressed Genes

A) For Gencode protein-coding genes without any overlapping annotated antisense transcript (N=10320), the average promoter methylation (TSS +/- 500bp) vs. the average ProSeq sense tag density in the gene body (TSS+1kb to TES) in mock treated cells. B) Average gene body methylation vs. average sense ProSeq tag density in mock treated cells. C) Average mock promoter methylation vs. average mock gene-body methylation. Most genes exhibit unmethylated promoters and methylated gene bodies. D) Fold changes (log2) in average ProSeq sense tag density upon DAC treatment among genes separated by percent promoter methylation in untreated cells (0-25%, 25-50%, 50-75%, or 75-100%). E) Fold changes (log2) in average ProSeq sense tag density upon DAC treatment among genes separated by percent gene-body methylation in untreated cells (0-25%, 25-50%, 50-75%, or 75-100%). F) Fold changes (log2) in average ProSeq sense tag density upon DAC treatment among genes separated by quantile of average ProSeq sense gene-body tag density in mock-treated cells. G) Average sense vs. antisense gene-body ProSeq tag density in mock-treated cells. H) Fold changes (log2) in average ProSeq antisense tag density upon DAC treatment among genes separated by percent gene-body methylation in untreated cells (0-25%, 25-50%, 50-75%, or 75-100%). I) Fold changes (log2) in average ProSeq antisense tag density upon DAC treatment among genes separated by quantile of average ProSeq antisense gene-body tag density in mock-treated cells. Only the genes with the most antisense transcription in mock-treated cells are sensitive to loss. Other genes tend to gain antisense tags. J) Homer was used call transcript peaks. Shown is the fraction of genes gaining or losing sense or antisense transcripts upon DAC treatment. K) Distribution of percent methylation in the promoter or gene-body in mock (green) or DAC (blue) treated cells, among those that gained or lost sense or antisense transcripts as called by Homer. L) Fold changes (log2) in gene-body tag density among genes significantly up- (N=1720) or down-regulated (N=1347) in breast cancer vs. normal tissue samples (TCGA, see text, Methods). M) Fraction of genes up- or down-regulated in breast cancer that are significantly up- or down-regulated by DAC treatment (FDR<.1).

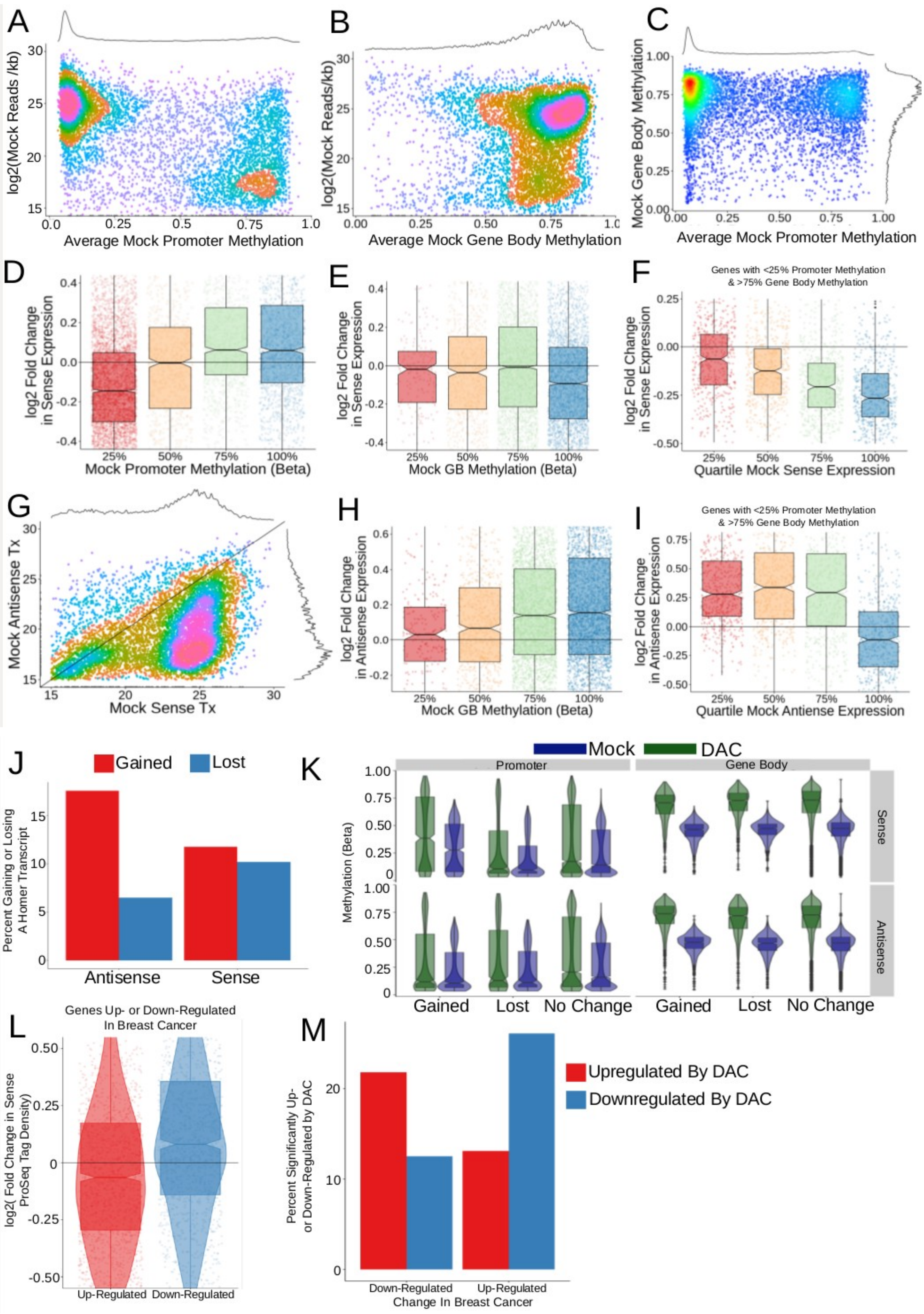

A
log2(Mock Reads /kb)
30
25
20
15
0.0
0.25
0.5
0.75
1.0
Average Mock Promoter Methylation
B
log2(Mock Reads/kb)
30
25
20
15
0.00
0.25
0.50
0.75
1.0
Average Mock Gene Body Methylation
C
Mock Gene Body Methylation
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Average Mock Promoter Methylation
D
log2 Fold Change in Sense Expression
0.4
0.2
0.0
-0.2
-0.4
25%
50%
75%
100%
Mock Promoter Methylation (Beta)
E
log2 Fold Change in Sense Expression
0.4
0.2
0.0
-0.2
-0.4
25%
50%
75%
100%
Mock GB Methylation (Beta)
F
Genes with <25% Promoter Methylation & >75% Gene Body Methylation
log2 Fold Change in Sense Expression
0.25
0.00
-0.25
-0.50
25%
50%
75%
100%
Quartile Mock Sense Expression
G
Mock Antisense Tx
30
25
20
15
15
20
25
30
Mock Sense Tx
H
log2 Fold Change in Antisense Expression
0.6
0.4
0.2
0.0
-0.2
25%
50%
75%
100%
Mock GB Methylation (Beta)
I
Genes with <25% Promoter Methylation & >75% Gene Body Methylation
log2 Fold Change in Antisense Expression
0.75
0.50
0.25
0.00
-0.25
-0.50
25%
50%
75%
100%
Quartile Mock Antiense Expression
J
Gained
Lost
Percent Gaining or Losing A Homer Transcript
15
10
5
0
Antisense
Sense
K
Mock
DAC
Promoter
Gene Body
Methylation (Beta)
1.00
0.75
0.50
0.25
0
Sense
Antisense
Gained
Lost
No Change
Gained
Lost
No Change
L
Genes Up- or Down-Regulated In Breast Cancer
log2(Fold Change in Sense ProSeq Tag Density)
0.50
0.25
0
-0.25
-0.50
Up-Regulated
Down-Regulated
M
Percent Significantly Up- or Down-Regulated by DAC
20
10
0
Upregulated By DAC
Downregulated By DAC
Down-Regulated
Up-Regulated
Change In Breast Cancer

Figure 5) Decitabine Results in Pervasive eRNA Induction and Preferentially Reactivates Enhancers Aberrantly Hypermethylated in Cancer

A) Enhancers were defined as overlapping H3K4me1/H3K27Ac peaks (see methods) more than 2kb from the nearest GenCode transcript. Enhancers meeting this definition in MCF7 cells were termed 'Active', while those active in other cell lines were termed 'Inactive'. ECGI are enhancers that overlap a UCSC-defined CGI, while Classical enhancers do not. Shown is the average DNA methylation in 20bp bins in mock-(green) and DAC-(blue) treated cells. B) Average normalized ProSeq tag density across each enhancer in mock-(green) and DAC-(blue) treated cells. C) Fold changes (log2) in ProSeq tag density in each set of enhancers upon DAC treatment. D) Scatterplot of average methylation in mock-treated cells vs. fold change in tag density (log2) upon treatment. Most enhancers are upregulated, especially if heavily methylated. E) Inactive enhancers were separated by those active in HMEC, a normal breast cell line. Shown are fold changes (log2) In average ProSeq tag density upon DAC treatment. Enhancers endogenously active in normal breast cells undergo greater activation, especially when hyper-methylated, than enhancers active in other cell types.

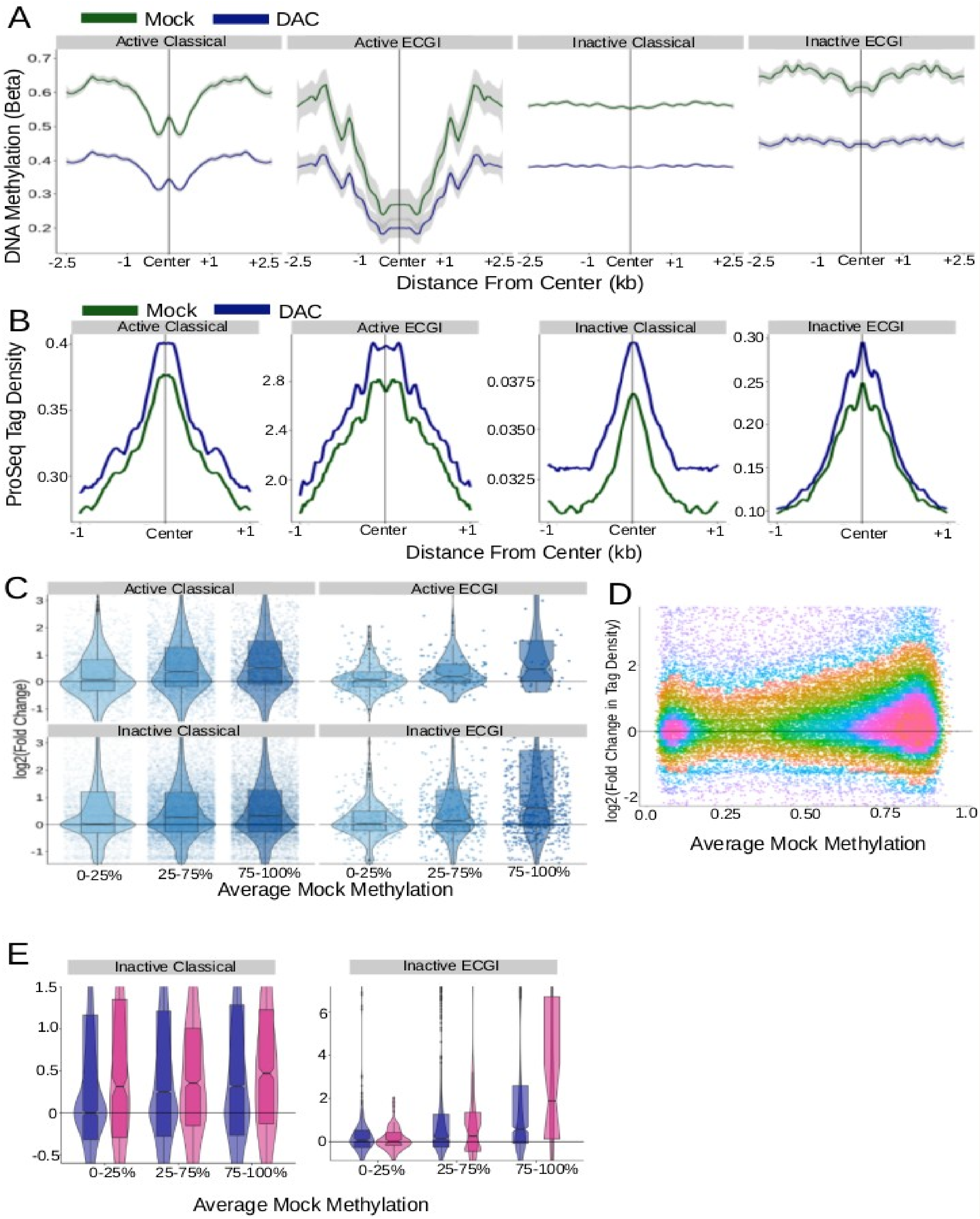
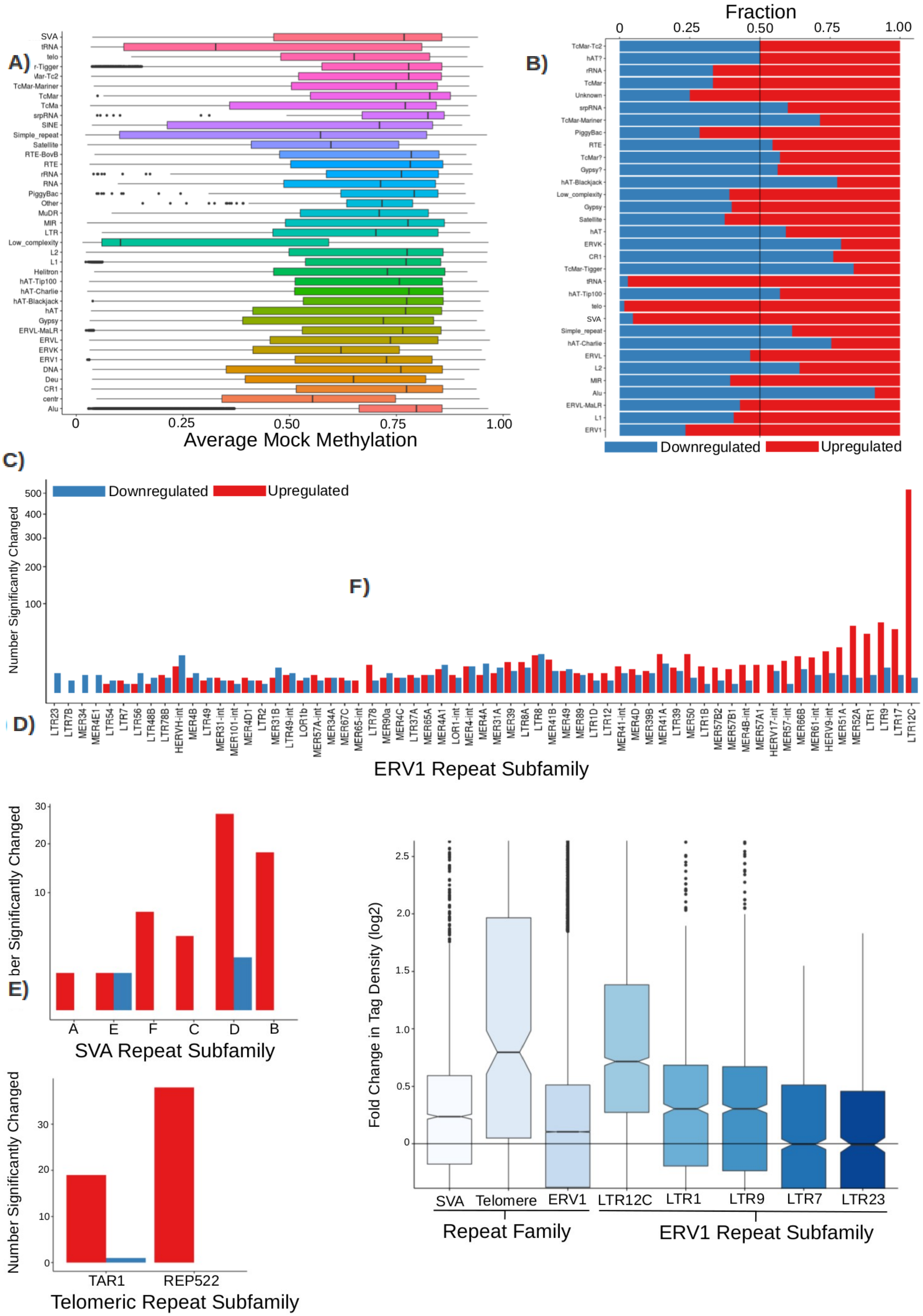

A
Mock
DAC
Active Classical
Active ECGI
Inactive Classical
Inactive ECGI
DNA Methylation (Beta)
0.7
0.6
0.5
0.4
0.3
0.2
-2.5
-1
Center
+1
+2.5
Distance From Center (kb)
B
Mock
DAC
Active Classical
Active ECGI
Inactive Classical
Inactive ECGI
ProSeq Tag Density
0.4
0.35
0.30
2.8
2.4
2.0
0.0375
0.0350
0.0325
0.30
0.25
0.20
0.15
0.10
-1
Center
+1
Distance From Center (kb)
C
Active Classical
Active ECGI
Inactive Classical
Inactive ECGI
log2(Fold Change)
3
2
1
0
-1
0-25%
25-75%
75-100%
Average Mock Methylation
D
log2(Fold Change in Tag Density)
2
0
-2
0.0
0.25
0.50
0.75
1.0
Average Mock Methylation
E
Inactive Classical
Inactive ECGI
1.5
1.0
0.5
0
-0.5
6
4
2
0
0-25%
25-75%
75-100%
Average Mock Methylation

Figure 6: Decitabine Unleashes Transcription of Repetitive Elements and ERVs

A) Feature annotated in the UCSC repeat database were grouped by repeat family. Shown in the distribution in average DNA methylation across each family. Most repeats are heavily methylated. B) Among repeats with uniquely mappable reads that changed significantly (FDR < 0.1) upon treatment, the fraction up- or down-regulated. Most repeat families are equally prone to gains or losses, except telomeres, SVA repeats, and ERV1, although many individual repeats in each family were induced. C) Number of features in each subfamily of the ERV1 family, D) SVA family, or E) telomeric repeat family significantly up- or down-regulated upon DAC-treatment. F) Distribution of fold changes (log2) in ProSeq tag density upon DAC-treatment in each repeat family (SVA, telomeric, ERV1), or select subfamilies of ERV1.


A)
SVA
tRNA
telo
r-Tigger
Mar-Tc2
TcMar-Mariner
TcMar
TcMa
srpRNA
SINE
Simple_repeat
Satellite
RTE-BovB
RTE
rRNA
RNA
PiggyBac
Other
MuDR
MIR
LTR
Low_complexity
L2
L1
Helitron
hAT-Tip100
hAT-Charlie
hAT-Blackjack
hAT
Gypsy
ERVL-MaLR
ERVL
ERVK
ERV1
DNA
Deu
CR1
centr
Alu
0
0.25
0.50
0.75
1.00
Average Mock Methylation
B)
Fraction
0
0.25
0.50
0.75
1.00
TcMar-Tc2
hAT?
rRNA
TcMar
Unknown
srpRNA
TcMar-Mariner
PiggyBac
RTE
TcMar?
Gypsy?
hAT-Blackjack
Low_complexity
Gypsy
Satellite
hAT
ERVK
CR1
TcMar-Tigger
tRNA
hAT-Tip100
telo
SVA
Simple_repeat
hAT-Charlie
ERVL
L2
MIR
Alu
ERVL-MaLR
L1
ERV1
🟦 Downregulated
🟥 Upregulated
C)
🟦 Downregulated
🟥 Upregulated
Number Significantly Changed
500
400
300
200
100
F)
LTR23
LTR7B
MER34
MER4E1
LTR54
LTR7
LTR56
LTR48B
LTR78B
HERVH-int
MER4B
LTR49
MER31-int
MER101-int
MER4D1
LTR2
MER31B
LTR49-int
LOR1b
MER57A-int
MER34A
MER67C
MER65-int
LTR78
MER90a
MER4C
LTR37A
MER65A
MER4A1
LOR1-int
MER4-int
MER4A
MER31A
MER39
LTR8A
LTR8
MER41B
MER49
MER89
LTR1D
LTR12
MER41-int
MER4D
MER39B
MER41A
LTR39
MER50
LTR1B
MER57B2
MER57B1
MER4B-int
MER57A1
HERV17-int
MER57-int
MER66B
MER61-int
HERV9-int
MER51A
MER52A
LTR1
LTR9
LTR17
LTR12C
D)
ERV1 Repeat Subfamily
ber Significantly Changed
30
20
10
E)
A
E
F
C
D
B
SVA Repeat Subfamily
Number Significantly Changed
30
20
10
0
TAR1
REP522
Telomeric Repeat Subfamily
Fold Change in Tag Density (log2)
2.5
2.0
1.5
1.0
0.5
0
SVA
Telomere
ERV1
LTR12C
LTR1
LTR9
LTR7
LTR23
Repeat Family
ERV1 Repeat Subfamily

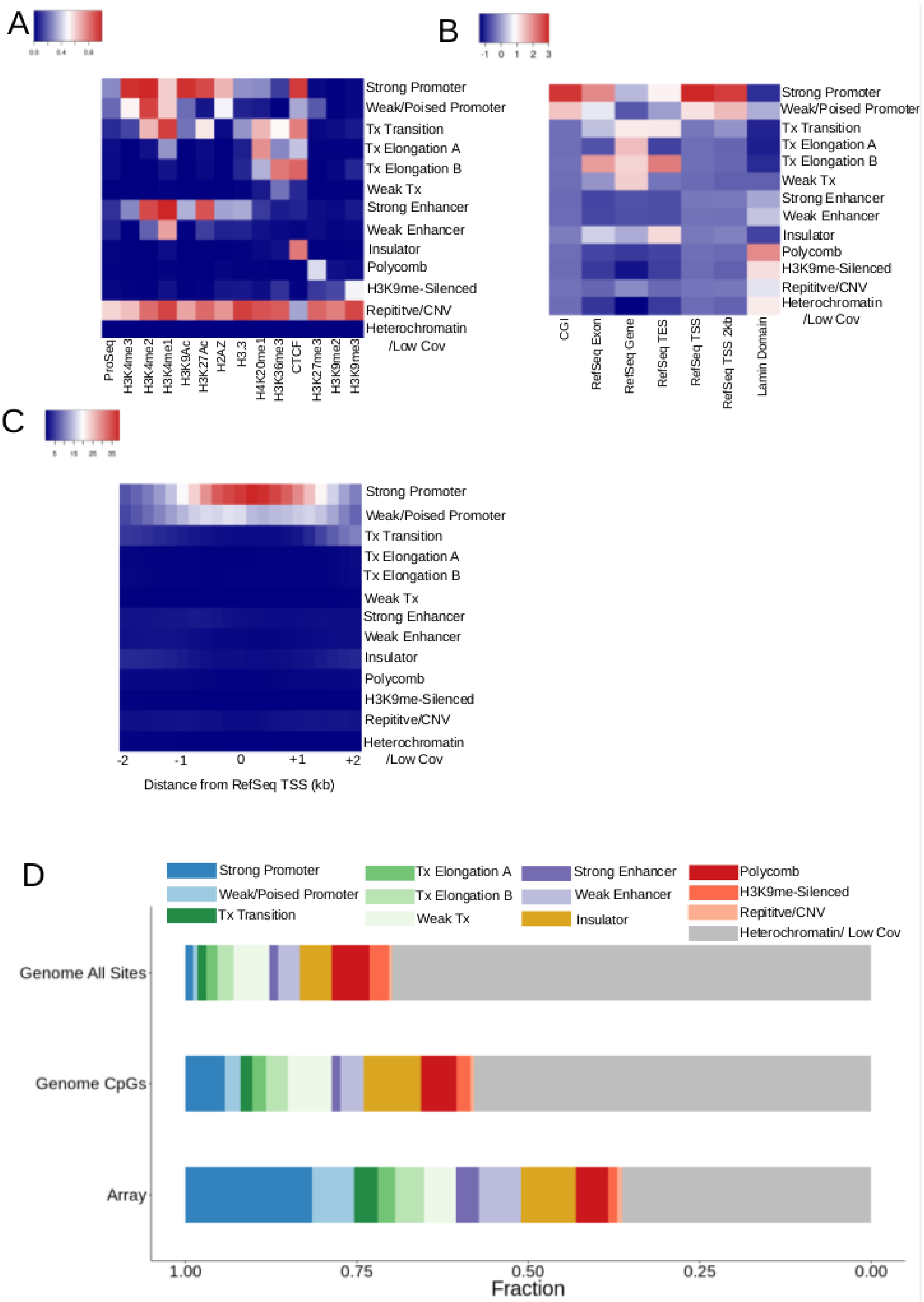
Supplemental Figure 1: ChromHMM for MCF7 cells

A) Enrichment of each ChIP-Seq (MCF7 ENCODE) or mock ProSeq in each chromatin state.

B) Enrichment in each genomic features for each chromatin state.

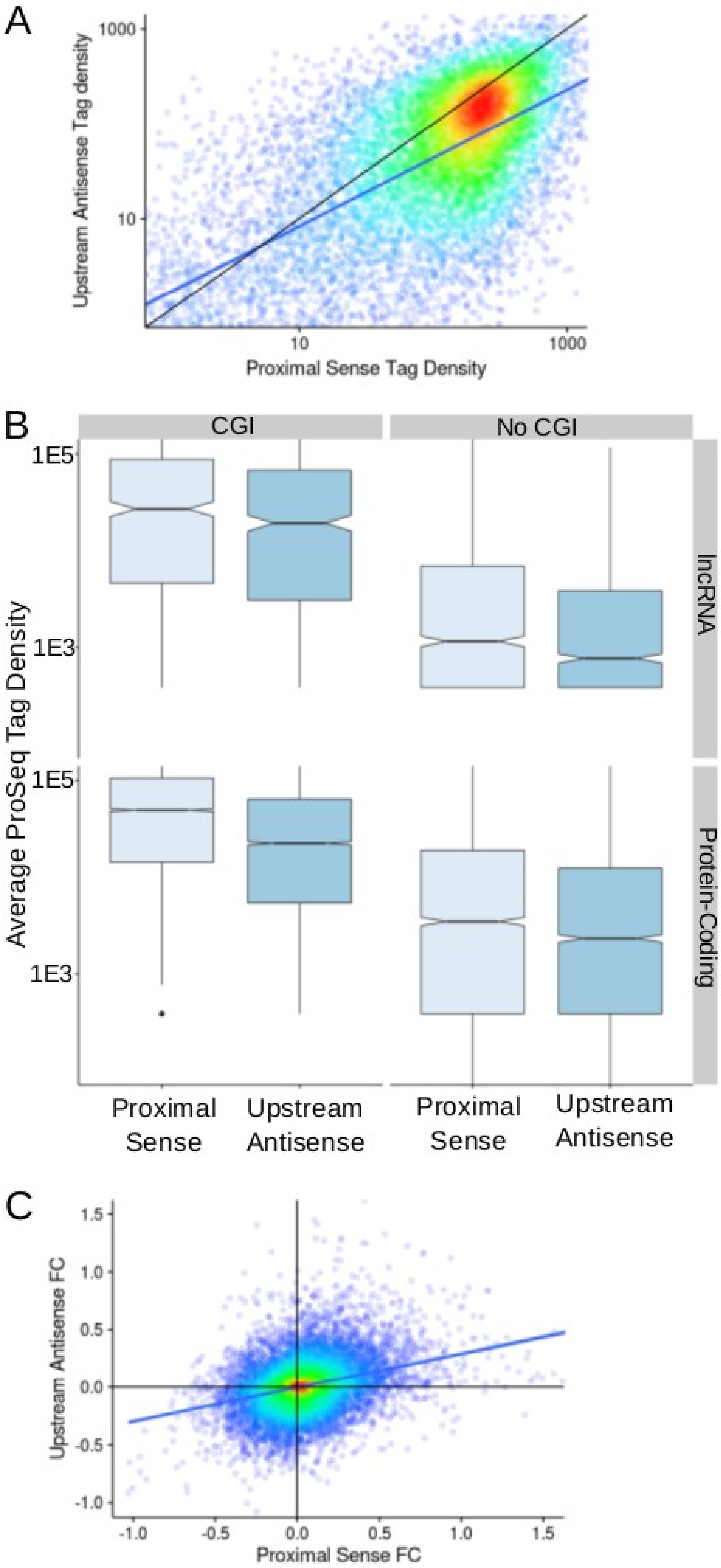
C) Enrichment in 200bp bins from annotated RefSeq TSSs for each chromatin state.

D) Distribution of the human genome, all CpGs in the genome, or CpGs covered by the EPIC array by chromatin state.


A
0.0
0.4
0.8
Strong Promoter
Weak/Poised Promoter
Tx Transition
Tx Elongation A
Tx Elongation B
Weak Tx
Strong Enhancer
Weak Enhancer
Insulator
Polycomb
H3K9me-Silenced
Repititve/CNV
Heterochromatin /Low Cov
ProSeq
H3K4me3
H3K4me2
H3K4me1
H3K9Ac
H3K27Ac
H2AZ
H3.3
H4K20me1
H3K36me3
CTCF
H3K27me3
H3K9me2
H3K9me3
B
-1
0
1
2
3
Strong Promoter
Weak/Poised Promoter
Tx Transition
Tx Elongation A
Tx Elongation B
Weak Tx
Strong Enhancer
Weak Enhancer
Insulator
Polycomb
H3K9me-Silenced
Repititve/CNV
Heterochromatin /Low Cov
CGI
RefSeq Exon
RefSeq Gene
RefSeq TES
RefSeq TSS
RefSeq TSS 2kb
Lamin Domain
C
5
15
25
35
Strong Promoter
Weak/Poised Promoter
Tx Transition
Tx Elongation A
Tx Elongation B
Weak Tx
Strong Enhancer
Weak Enhancer
Insulator
Polycomb
H3K9me-Silenced
Repititve/CNV
Heterochromatin /Low Cov
-2
-1
0
+1
+2
Distance from RefSeq TSS (kb)
D
Strong Promoter
Weak/Poised Promoter
Tx Transition
Tx Elongation A
Tx Elongation B
Weak Tx
Strong Enhancer
Weak Enhancer
Insulator
Polycomb
H3K9me-Silenced
Repititve/CNV
Heterochromatin/ Low Cov
Genome All Sites
Genome CpGs
Array
1.00
0.75
0.50
0.25
0.00
Fraction

Supplemental Figure 2: DAC treatment upregulates upstream antisense transcription, which is tied to initiation at the paired sense promoter.

A) Among all protein-coding Gencode genes, the average ProSeq tag density in mock-treated cells at the sense promoter (TSS +250bp, only sense tags) or upstream antisense transcript (TSS -250bp, only antisense tags). B) Distribution of average sense and antisense promoter transcripts among protein-coding genes and Gencode lncRNAs, among those where the TSS overlaps a CGI, or those at least 2kb from a CGI. C) Fold change (log2) in tag density upon DAC treatment at paired sense and antisense promoters.


A
Upstream Antisense Tag density
1000
10
Proximal Sense Tag Density
10
1000
B
CGI
No CGI
lncRNA
Protein-Coding
Average ProSeq Tag Density
1E5
1E3
1E5
1E3
Proximal Sense
Upstream Antisense
Proximal Sense
Upstream Antisense
C
Upstream Antisense FC
1.5
1.0
0.5
0.0
-0.5
-1.0
Proximal Sense FC
-1.0
-0.5
0.0
0.5
1.0
1.5

Bibliography

1. Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26,** 577–590 (2014).
2. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
3. Kellner, W. A., Bell, J. S. & Vertino, P. M. GC skew defines distinct RNA polymerase pause sites in CpG island promoters. *Genome Res.* **25,** 1600–1609 (2015).
4. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13,** 720–731 (2012).
5. Ding, X.-L., Yang, X., Liang, G. & Wang, K. Isoform switching and exon skipping induced by the DNA methylation inhibitor 5-Aza-2′-deoxycytidine. *Sci. Rep.* **6,** (2016).
6. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).
7. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–220 (2013).
8. Aran, D. & Hellman, A. DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* **154,** 11–13 (2013).
9. Barwick, B. G., Scharer, C. D., Bally, A. P. & Boss, J. M. Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation. *Nat. Immunol.* **17,** 1216–1225 (2016).
10. Bell, R. E. *et al.* Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.* gr. 197194.115 (2016).
11. Chiappinelli, K. B. *et al.* Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162,** 974–986 (2015).
12. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11,** 1455–1476 (2016).

13. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17,** 208 (2016).

14. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).

15. Consortium, E. P. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306,** 636–640 (2004).

16. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12,** 433–438 (2015).

17. Rosenbloom, K. R. *et al.* The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* **43,** D670–D681 (2015).

18. Liu, M. *et al.* Vitamin C increases viral mimicry induced by 5-aza-2′-deoxycytidine. *Proc. Natl. Acad. Sci.* **113,** 10238–10244 (2016).

19. Rao, X. *et al.* CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene* **32,** 4519–4528 (2013).

20. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41,** 178–186 (2009).

21. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150,** 12–27 (2012).

22. Lee, J.-H. & Skalnik, D. G. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.* **280,** 41725–41731 (2005).

23. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12,** 283–293 (2011).

24. Yu, W. *et al.* Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways. *Genome Res.* **24,** 1613–1623 (2014).

25. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16,** 519–532 (2015).

26. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543,** 72–77 (2017).

27. Bell, J. S. K. *et al.* Factors affecting the persistence of drug-induced reprogramming of the cancer methylome. *Epigenetics* **11,** 273–287 (2016).

28. Venkatesh, S., Li, H., Gogol, M. M. & Workman, J. L. Selective suppression of antisense transcription by Set2-mediated H3K36 methylation. *Nat. Commun.* **7,** (2016).

29. Gordon, A. & Hannon, G. J. Fastx-toolkit. *FASTQA Short-Reads Preprocessing Tools Unpubl. Httphannonlab Cshl Edufastxtoolkit* (2010).

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

31. Wysoker, A., Tibbetts, K. & Fennell, T. Picard tools version 1.90. *http://picard.sourceforge.net* (2013).

32. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9,** e1003118 (2013).

33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

34. Morris, T. J. *et al.* ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics* **30,** 428–430 (2014).

35. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).

36. Zhao, S., Guo, Y., Sheng, Q. & Shyr, Y. Advanced heat map and clustering analysis using heatmap3. *BioMed Res. Int.* **2014,** (2014).

37. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).

Chapter VI: Discussion

This work has made a number of fundamental contributions towards the understanding the relationship between DNA methylation and transcription, two of the most basic nuclear processes with critical roles in human disease and development, while discovering new mechanisms by which epigenetic therapy reprograms the cancer transcriptome.

CpG Islands are Unique Chromatin Environments

CpG Islands are evolutionary evidence of a genomic region's selection to remain unmethylated, and thus are thought to represent biologically important sites, yet many questions remain surrounding their function. The canonical role of CGI has been as strong promoters [1]. Comparing CGI to other promoters, reveals that CGI have much higher levels of divergent transcription, H3K4me3, H3K27Ac, and any other measure of promoter strength compared to CGI-poor promoters. Given these high transcriptional initiation levels, we addressed the hypothesis that CGI may have unique transcriptional profiles as well, by evaluating the relationship between transcription (GroSeq) and CGI. We found a striking accumulation of nascent transcription at CGI boundaries, suggesting a novel regulatory step. We determined that genes tend to primarily exhibit proximal or distal pausing, although most exhibited both to at least some extent. We were able to link this novel transcriptional regulatory step at promoters wherein Pol II pauses at CGI boundaries to GC-skew, which we found dictated the degree and location of both proximal and distal pausing. Recent work has focused on the role of R-loops (which are most stable in regions of high GC-skew) in preventing DNA methylation in promoters[2], but also at transcription end sites, where they apparently help arrest transcription[3]. This work suggests that R-loops in fact play both roles at

CGI promoters-preventing DNA methylation and slowing transcriptional elongation- helping to unify these phenomena. We also noted that upstream antisense transcription appears to proceed at least to the 5' CGI boundary. When we examined the remethylation of CGI, this nascent transcription appeared to protect CGI from *de novo* methylation. Even CGI that had apparently become methylated in MB231 cells were able to repel new methylation once it was lost to DAC treatment, implicating Pol II occupancy in protecting the hypomethylation of CGI. Given the tight association with Pol II initiation and histone modifications like H3K4me3 and H3K9/27Ac, it can be difficult to assign a role in preventing DNA methylation directly to transcription in this regard, but the fact that transcription extends far past the TSS of these genes to the CGI boundaries, suggests an active role for transcription per se.

There is substantial evidence that enhancers and promoters share a common architecture [4], and differ primarily in simple transcript stability. Our findings place ECGI on a continuum of Pol II initiation potential somewhere between classical CpG-poor enhancers and promoter CGI, which have higher levels of all the aforementioned correlates of initiation potential and enhancer/promoter strength. The identification of distal pausing and resistance to *de novo* methylation at CGI helped highlight the unique nature of CGI as sites for massive transcriptional initiation and apparent sequence-based resistance to DNA methylation that contribute to their strength as promoters. Yet, some half of CGI do not in fact contain an annotated TSS, and have been referred to as 'orphan' CGI. The field has long assumed these sites to be simple unannotated promoters [4]. Yet RNA-Seq across a plethora of cell lines and conditions in recent years has failed to demonstrate stable transcripts originating from these sites, even as techniques developed to look for actual Pol II initiation (GroSeq, ChIP-Seq) hinted that intergenic CGI were transcriptionally active. Using data designed to discriminate stable from unstable from stable transcripts (GroCap vs. CAGE) we were able to determine that many CGI, even those within or near genes, were likely to give rise to strictly unstable transcripts, which are the hallmark of eRNAs. While most CGI near or within genes were also likely to contain stable transcripts, implicating them as possible alternative promoters, orphan CGI nearly always contained unstable transcript pairs. By analyzing chromatin profiles across over a hundred primary cell types and cell lines, we documented that the vast majority of these orphan CGI had an active enhancer chromatin state (H3K4me1/H3K27Ac) in at least one lineage. Mirroring the transcript stability findings, a large portion of perigenic, and even TSS-containing promoter CGI, were likely to manifest an enhancer-like state. In screens of enhancer activity, we found that CGI in general, even those with a TSS, were able to strongly drive transcription of a reporter gene. Together, these data suggest that enhancer activity may be an intrinsic feature of all CGI, including promoter CGI. This finding is critical to understanding why CGI are such strong promoters when unmethylated and, in addition, suggests a role for not only the thousands of intragenic and perigenic CGI, but especially orphan CGI as enhancers, or ECGI.

Furthermore, the same features that appear to empower promoter CGI also engender ECGI with tremendous strength relative to classical or CpG-poor enhancers: CpG density, preserved by hypomethylation, ensures the binding of GC and CpG-binding transcription factors, enables methylation-sensitive binding of CTCF and proper chromatin topology, recruitment of active chromatin modifiers, and ultimately transcriptional initiation. These features lead to much higher H3K4me3/H3K4me1 ratios at ECGI relative to classical enhancers, a common gauge of enhancer strength. Virtually every other indicator of enhancer activity: H3K9/27Ac, DNAse hypersensitivity, transcription factor binding, and genomic contacts, are also higher at ECGI than classical enhancers, and are features shared by canonical promoter CGI. Indeed, many of the same TFs enriched at ECGI are also enriched at promoter CGI, in particular those that bind GC and CpG-rich DNA.

CGI Conservation

A large body of work also supports the idea that enhancers in general are less conserved than promoters [6], which has been interpreted to reflect their roles as adaptable accessories to genes, rather than intrinsic elements of genes critical to their function. However, the extensive conservation of ECGI within mammals, even extending to distant non-placental marsupials (opossum) and even more distant monotremes (platypus) suggests that ECGI may have played critical roles in mammalian evolution. While ECGI are relatively few in number, the immense strength they possess as enhancers, as well as important roles in chromatin topology (as CTCF binding sites and TAD-neighbors), may have enabled them to influence gene expression in unexpected ways, for example by altering broad chromatin domains through position effects or causing a gene to be expressed across a wider range of lineages.

In evolutionary terms, it has been shown that CGI promoters exhibit conservation of their CpG sites, but not other residues [7]. This suggests that CpG density of such promoters is the primary factor in their ability to act as promoters, especially as many promoter CGI lack TATA boxes and other sequence-based features typically ascribed to promoters. We found that ECGI exhibited the same trend, with their CpG residues extensively conserved among mammals, but not among vertebrates as a whole. Promoter CGI exhibited even greater CpG conservation among mammals, and more broadly among vertebrates. This pattern was evident not only in the conservation scores for individual residues, but also in the retention of CGI status in various other animals. Thus, the fact that ECGI exhibit less H3K4me3 and other features of activity than promoter CGI (albeit much more than classical CpG-poor enhancers), appears to translate directly to their ability to repel mutagenic DNA methylation over evolutionary time. Among promoter CGI, the vast majority are hypomethylated in embryonic stem cells. Because only mutations in the germline are transmitted to offspring, methylation in somatic tissues does not affect CpG density over evolutionary time, but does have important implications for regulation and studies of sensitivity to *de novo* methylation. We found a similar trend among ECGI regarding embyronic stem cell methylation, but not to the same extent as promoter CGI, again reflecting the overall lower conservation and euchromatic features of ECGI compared to promoter CGI.

A minority of promoter CGI are known to undergo programmed silencing during the course of development, and this often coincides with hypermethylation. However, there has been longstanding debate in the field as to whether hypermethylation is directly responsible for gene silencing, or merely a consequence of inactivity and the lack of the transcription and active histone modifications or R-loops necessary to repel *de novo* methylation. Substantial evidence suggests that most promoter CGI that undergo hypermethylation are first silenced by the Polycomb complex and concomitant H3K27me3[8]. Indeed, accumulating evidence suggests that CGI are inherently able to recruit polycomb [9], and that CGI serve as the mammalian equivalent of *Drosophila* Polycomb Recruitment Elements (PRE), which are also GC and CpG-rich even though *Drosophila* lacks DNA methylation, and thus CGI. Such elements were also found to be high in H3K4me2 when active, just like active CGI (especially ECGI which possess higher levels of H3K4me2 than promoter CGI), suggesting overlap between these elements.

Polycomb occupancy is not sufficient for silencing, though, and many CGI and other regions exhibit so-call bivalency, that is, occupancy of both repressive marks like H3K27me3, but also active marks like H3K4me3 and H3K27Ac (necessarily on histones lacking

H3K27me3). Indeed, we find that ECGI often possess H3K27me3 even when active, but especially in cell lineages in which they are inactive. Like at promoter CGI, this may implicate H3K27me3 as a marker for future DNA methylation, which is thought to be a more stable silencing mechanism than histone-based means. Indeed, we find evidence that a substantial fraction of ECGI undergo DNA methylation in normal somatic tissues, at a greater rate than promoter CGI. However, polycomb recruitment may serve a broader role at ECGI in particular to silence broader domains and not just the ECGI itself. Further evidence for this model comes from the fact that ECGI, like promoter CGI, but unlike classical enhancers, are enriched at TAD boundaries, which demarcate active from inactive chromatin domains. Thus, ECGI may function as a global mechanism by which polycomb domains are established during development, or perhaps tumorigenesis.

CGI associated with pseudogenes are less conserved, and likely in the process of losing their CpG density as their selection to remain unmethylated fades with the coding potential of their associated genes. Indeed, they tend to be heavily methylated, and we observed very little evidence of TSSs, stable or unstable, at such sites. We also found orphan CGI with similar heterochromatic features and lack of transcriptional initiation among Remnant CGI, those orphan CGI for which we were unable to identify enhancer chromatin features in any cell line examined. These Remnants were less conserved than ECGI even in closely related mammals, and exhibited methylation in embryonic stem cells, a hallmark of mutagenicity in CpG sites. While it is possible that some of the Remnants were once associated with pseudogenes, the lack of annotated open reading frames at these loci suggest that they are instead former ECGI. Their identification thus not only affirms the evolutionary principle that selection to remain unmethylated is required for the preservation of CpG density, but also highlights the critical function of euchromatic CGI, both as promoters and as enhancers.

DNA Methylation is a Master Regulator of Transcriptional Initiation

Further evidence for the unique role of CGI of all stripes come from our final study on how transcriptional dynamics change upon DAC-induced demethylation. Consistent with the initial hypomethylated status, active promoter CGI exhibit only modest changes upon DAC treatment as they have little methylation to lose. That said, they do gain a statistically significant amount of new of Pol II initiation, suggesting that methylation may play a role in buffering initiation levels at nearly all CGI, if a modest one. Notably, most of this newly engaged Pol II is incapable of escaping past the proximal and distal pause points, and CGI genes do not gain more gene body transcripts on the whole. Silent CGI with DNA hypermethylation, however, are prone to robust increases in initiation, at least as gauged by Pol II occupancy in the promoter. However, demethylation of CGI promoters is not sufficient to yield nearly as much initiation as seen as endogenously active CGI promoters. This suggests that methylation is indeed a silencing mechanism at these loci, but that additional features like TF binding or histone modifications may be required for full activity.

Nevertheless, these data confirm that CGI, even in a heterochromatic environment, apparently have an intrinsic, sequence-based proclivity to transcriptional initiation that is lacked, or at least not shared to the same extent, by CpG-poor promoters. In addition to apparently requiring further factors for full initiation potential, both proximal and distal pausing remain intact at silenced CGI: we observe dramatic decreases in Pol II activity immediately after the TSS where promoter-proximal pausing occurs, and further decreases after the CGI boundary where distal pausing occurs. This implicates sequence based features, particularly the GC-skew we previously implicated in both pausing sites, as essential regulators of transcription.

ECGI show a similar relationship between methylation and initiation, with those that lose substantial methylation gaining substantial transcription. These effects are greater than those seen at classical enhancers, perhaps because of the increased CpG density of ECGI.

As discussed above, CGI also possess abundant upstream antisense transcription that may help them retain their open chromatin state [4,14]. Mirroring the modest increases in sense initiation, active unmethylated CGI also acquired slightly more upstream antisense transcription, while silent methylated CGI exhibit similar activation of uaRNA and mRNA, suggesting that methylation serves to inhibit initiation at both paired TSSs. We identified the same phenomenon at ECGI, which also possess paired eRNA transcript start sites. Notably, CGI associated with known pseudogenes also became activated by DAC-induced demethylation, but again these gains in initiation were limited to the CGI by promoter-proximal and distal pausing, and initiation levels did not approach those of active CGI associated with intact genes.

Thus these studies have provided a great deal of new insight into the evolutionary history and function of CpG Islands. We have established that they are unique transcription units with a capacity to limit transcription past their boundaries, that they are extremely resilient to *de novo* methylation, that they can function as powerful enhancers, and that their activity can be buffered by DNA methylation, even though other factors are required for full transcriptional potential.

While CGI are attractive models for studying DNA methylation, especially given their unique hypomethylated state and critical roles in genome biology, they represent a comparatively minor portion of the genome as a whole which represents a largely hypermethylated landscape. Pol II is thought to be a somewhat promiscuous complex, binding to accessible chromatin and sampling it for transcriptional potential [15]. Thus, DNA elements with initiation potential throughout the genome that are outside of the purview of a given cell lineage's transcriptional program must be prevented from attracting Pol II if spurious transcripts and titration of Pol II from desired targets are to be achieved. DNA methylation is generally thought to mask such spurious TSS, but this concept is difficult to test using RNA-Seq, which can only detect stable transcripts of a certain concentration. Thus, we performed the first genome-wide experiment designed to assay nascent, unstable transcripts (utilizing ProSeq) while modulating DNA methylation levels to test this model genome-wide. Indeed, we found extensive evidence that not only are spurious transcripts unmasked by DAC, but methylation also controls the transcriptional levels at non-CGI promoters, classical enhancers, pseudogenes, and many repetitive elements.

Among non-CGI promoters, we found that most exhibit fairly high methylation to begin with. Among those that are endogenously unmethylated, we observed much higher expression than among methylated promoters. Consistent with this, we observed that methylation loss directly correlated with *de novo* initiation upon DAC-induced demethylation. However, like at CGI promoters, most of this Pol II was unable to escape Pol II pausing and overall genes with methylated promoters exhibited very modest gains in gene body transcription, a measure of mRNA production, relative to their gains in initiation. Such promoters also exhibited modest gains in upstream antisense transcription, again consistent with CGI promoters. This relationship with methylation at non-CGI promoters has been poorly studied until now because initiation/pausing dynamics can not be captured by traditional RNA-Seq. Consistent with lower transcription, we also observed that non-CGI TSS are much more prone to regaining methylation after DAC treatment than are CGI promoters. This may also be a reflection of their diminished H3K4me3 and H3K9/27Ac.

DNA methylation has also be linked to enhancer activity, with active enhancers typically displaying lower methylation levels than inactive enhancers [16,17], but no studies thus far have assayed eRNA production genome-wide while altering DNA methylation levels. With ProSeq, we were able to document massive induction of eRNA across both enhancers already active in MCF7 and hundreds of thousands of others that would ordinarily be restricted to other lineages. Classical enhancers, even when active, still tend to exhibit modest transcriptional levels which do not approach those seen at CpG-poor promoters with similar methylation levels, possibly indicating that they have less sequence-based transcriptional potential. Consistent with a role of transcription and its associated histone environment in preventing DNA methylation, we observed that following DAC-induced demethylation, strong enhancers (those with greater histone acetylation levels as defined by HMM) were much more resilient to remethylation than weak enhancers.

Enhancers and promoters are not the only genomic elements with proclivities towards transcriptional initiation, and while lineage-specific methylation ensures fidelity to the transcriptional programs at these needed transcripts, DNA methylation is also vital to repressing DNA repeats, especially endogenous retroviruses [19,20]. Consistent with this, we found that DAC induced several repeat families including pericentromeres and telomeres, and especially the LTR12C ERV. Among pseudogenes, demethylation appeared sufficient to induce some level of initiation, but the additional barrier of Pol II pausing prevented most, though not all, pseudogenes from gaining appreciable gene body transcription. The consequences of pseudogene activation are unclear, but many are likely to produce short dysfunctional proteins, or at the very least occupy splicing machinery, ribosomes, and proteosome components that would otherwise be devoted to the proper transcriptional program. Pseudogene proteins, depending on their specific nature, may also play dominant negative roles in cells [22] by oligomerizing or forming complexes with wild-type proteins to impair their function, or perhaps by ineffectually occupying TF binding sites.

We also observed abundant antisense transcription within genes, which was largely upregulated by demethylation. This is consistent with the unmasking of spurious TSSs within the gene body that are normally silenced by DNA methylation. It is unclear if this antisense transcription interferes with elongation of the sense transcript, as generally levels of sense and antisense transcription were positively correlated. In general, DNA methylation levels in gene bodies are associated with higher sense expression. We observe this pattern in our ProSeq data, but this correlation is quite modest and most genes exhibit heavy DNA methylation even if lowly transcribed. Importantly, though, we observed that higher expression levels sensitized genes to more downregulation by DAC treatment.

Resetting the Cancer Epigenome

Epigenetic drugs are a promising avenue for cancer therapy, with several drugs currently FDA-approved, and many more in clinical trials and preclinical development. The primary mechanism of epigenetic therapy is thought to be resetting the epigenetic state of the cancer genome to that of a more normal genome, and thus more normal phenotype. We found effects of DAC treatment that support this model: a reliance of over-expressed oncogenes on gene body methylation, reactivation of genes silenced by promoter methylation, and up-regulation of enhancers silenced during cancer progression (Chapter V). Furthermore, we were able to show that DAC-induced reversal of endogenous methylation returns quickly, but that aberrant cancer-related hypermethylation is unlikely to return (Chapter III). Indeed, in that study we were able to show that promoters and CGI in particular were resistant to remethylation, as were strong enhancers (more highly transcribed) compared to weak enhancers. Thus, not only is aberrant methylation more likely to be lost during DAC treatment, it tends to be forgotten while endogenous methylation returns rapidly, validating the continued interest in epigenetic therapy.

Future Directions

As any quality research, our work here has produced as many questions as it has answered. First, the full scope of genomic regulation by GC-skew, and the mechanisms of the enforcement and release of Pol II pausing remain to be studied. While we established a clear role of GC-skew in dictating the position and degree of Pol II pausing in CGI promoters, we have not undertaken a comprehensive analysis of the effects of GC-skew within gene bodies or at exon/intron boundaries where Pol II elongation also slows. Enhancers, or ECGI, may also utilize GC-skew to prevent further elongation of eRNA transcripts. Likewise, it is possible that GC-skew at spurious transcripts or in endogenous retroviruses or repeats could serve as another safeguard against expression. It would also be interesting to study mutation in regions of GC-skew in cancer cells to determine if loss of guanines on the coding strand could be a mechanism to permit oncogene over-expression, or if gain of guanines could reduce expression of tumor-suppressor genes, although this question awaits larger number of whole genome sequences of cancer cells. Similarly, germline SNPs may contribute or detract from GC-skew to affect gene expression levels across humans. While R-loops may play an important role in mediating Pol II pausing at regions of GC-skew, the mechanism of release from such pausing is unclear. It is possible that helicases or other factors are necessary to resolve the R-loop, but even their identification would yield questions about their mechanisms of recruitment. Given that enhancers appear to contact the dominant pausing site, such factors may be recruited by enhancers; this would have important implications for ECGI in particular given that they, like promoter CGI, manifest R-loops, and thus ECGI may be primed with such factors to deliver to the target gene promoter.

While we directly assayed how DNA methylation affects nascent transcription, it would be intriguing to examine how histone modifications individually contribute to nascent transcription, as well as how histones are changed by DAC treatment. Indeed, decitabine has been shown to result in widespread retargeting of polycomb and H3K9me3 [30], which largely silence distinct chromatin domains, and it would be interesting to observe how HDAC inhibition combines with DNA demethylation in terms of Pol II initiation, release from pausing, and potentiating the activation of repeats and cryptic promoters. It may also be interesting to examine nascent transcription changes with a combination of decitabine and inhibition of the histone demethylase LSD1, which shows synergistic effects with DAC in inhibiting AML cell growth and survival [31]. The accumulation of H3K4me3 may license even greater reactivation of hypermethylated CGI in particular, given their predisposition to broad H3K4me3 recruitment.

H3K36me3 levels correlate with DNA methylation in gene bodies, and is thought to play similar roles in prevention aberrant initiation. Analyzing nascent transcription with the inhibition of both marks with appropriate controls may help uncover the redundancy or specificity of each. The full extent of how chromatin marks influence remethylation kinetics following DAC-induced demethylation also need to be directly addressed. While we provided correlative evidence, both negative and positive, for a number of marks, mapping each in a similar time-course experiment via ChIP-Seq would help answer a number of basic epigenetic questions about the temporal and spatial relationships among various epigenetic marks, whether H3K36me3 recruits DNA methylation or vice versa, if DNA methylation proceeds

H3K27me3 as it appears to during development, and the extent to which H3K4me is allowed to accumulate in newly demethylated promoters, for example.

Finally, the identification of ECGI as a class of highly active enhancers raises a multitude of questions about their roles in development, mammalian evolution, nuclear organization, polycomb recruitment, and cancer. While we found that ECGI were more broadly expressed across cell types, ECGI with highly lineage-restricted expression may play interesting roles driving lineage-defining genes. Given the rapid evolution of ECGI compared to promoter CGI, neuronal lineages (which may have evolved rapidly in humans, and possess high 5hmC levels common at ECGI) in particular may warrant further investigation as more epigenomic data for such cell types becomes available. With the advent of Hi-C, we are just beginning to be able to assay chromatin contacts genome-wide in high resolution. This will present unique opportunities for correlating epigenetic changes in ECGI with changes in chromatin topology, and indeed assaying chromatin architecture with Hi-C following decitabine treatment would answer a number of fundamental questions about how DNA methylation affects CTCF binding to dictate TAD formation. Given their substantial activity, it would be interesting to compare enhancer methylation changes across cancer types to determine how cancer cells are selected to aberrantly silence ECGI or co-opt them from other lineages.

Our identification of enhancer methylation as a therapeutic target in cancer suggests that additional activation of enhancers may be useful. For example, DAC has shown some success with the HDAC inhibitor entinostat in breast cancer patients [32], and with Polycomb inhibition in leukemia cells[33]. These combination therapies likely give greater induction to enhancers, as well as repetitive elements and spurious promoters activated by DAC, which can be further silenced by lack of histone acetylation or H3K27me3. Epigenetic therapy is one of the most promising avenues of cancer research, and this work has significantly advanced our understanding of its effects, as well as the changes in methylation that help drive oncogenesis.

Bibliography

1. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
2. Hartono, S. R., Korf, I. F. & Chédin, F. GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. *Nucleic Acids Res.* gkv811 (2015).
3. Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I. & Chédin, F. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.* **23,** 1590–1600 (2013).
4. Lis, J. *et al.* A Unified Model Describing The Architecture And Creation Of Promoters And Enhancers. *FASEB J.* **29,** 497.3 (2015).
5. Illingworth, R. S. *et al.* Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* **6,** e1001134 (2010).
6. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).
7. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38,** 626–635 (2006).
8. Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* **1,** e1500737 (2015).
9. Rickels, R. *et al.* An Evolutionary Conserved Epigenetic Mark of Polycomb Response Elements Implemented by Trx/MLL/COMPASS. *Mol. Cell* **63,** 318–328 (2016).
10. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150,** 12–27 (2012).
11. Lee, J.-H. & Skalnik, D. G. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast

Set1/COMPASS complex. *J. Biol. Chem.* **280,** 41725–41731 (2005).

12. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–357 (2012).
13. Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13,** 1924–1935 (1999).
14. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).
15. Faralli, H. & Dilworth, F. J. Chaperoning RNA Polymerase II through repressive chromatin. *EMBO J.* **32,** 1067–1068 (2013).
16. Aran, D. & Hellman, A. DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* **154,** 11–13 (2013).
17. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12,** 283–293 (2011).
18. Liu, W. *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155,** 1581–1595 (2013).
19. Yu, W. *et al.* Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways. *Genome Res.* **24,** 1613–1623 (2014).
20. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16,** 519–532 (2015).
21. Chiappinelli, K. B. *et al.* Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162,** 974–986 (2015).
22. Faial, T. BRAF pseudogene induces cancer. *Nat. Genet.* **47,** 429–429 (2015).
23. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543,** 72–77 (2017).
24. Venkatesh, S., Li, H., Gogol, M. M. & Workman, J. L. Selective suppression of antisense transcription by Set2-mediated H3K36 methylation. *Nat. Commun.* **7,** (2016).
25. Ding, X.-L., Yang, X., Liang, G. & Wang, K. Isoform switching and exon skipping induced by the

DNA methylation inhibitor 5-Aza-2′-deoxycytidine. *Sci. Rep.* **6,** (2016).

26. Tsai, H.-C. *et al.* Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell* **21,** 430–446 (2012).
27. Theodorou, V., Stark, R., Menon, S. & Carroll, J. S. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res.* **23,** 12–22 (2013).
28. Bell, R. E. *et al.* Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.* gr. 197194.115 (2016).
29. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* **16,** 483–493 (2016).
30. Komashko, V. M. & Farnham, P. J. 5-azacytidine treatment reorganizes genomic histone modification patterns. *Epigenetics* **5,** 229–240 (2010).
31. Duy, C. *et al. Cooperative Gene Repression By DNA Methylation and LSD1-Mediated Enhancer Inactivation in Acute Myeloid Leukemia*. (Am Soc Hematology, 2016).
32. Connolly, R. *et al.* Combination Epigenetic Therapy in Advanced Breast Cancer with 5-Azacitidine and Entinostat: A Phase II National Cancer Institute/Stand Up to Cancer Study. *Clin. Cancer Res.* clincanres. 1729.2016 (2016).
33. Momparler, R. L., Idaghdour, Y., Marquez, V. E. & Momparler, L. F. Synergistic antileukemic action of a combination of inhibitors of DNA methylation and histone methylation. *Leuk. Res.* **36,** 1049–1054 (2012).