Title
Determinants of Unique DNA Methylation, Histone Modification, and Nucleosome Occupany at CpG Islands

Author
Langerman, Justin Bryon

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Determinants of Unique DNA Methylation, Histone Modification, and Nucleosome Occupany at CpG Islands

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Molecular Biology

by

Justin Bryon Langerman

2014

ASTRACT OF THE DISSERTATION

Determinants of Unique DNA Methylation, Histone Modification, and Nucleosome Occupany at CpG Islands

by

Justin Bryon Langerman

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2014

Professor Stephen T. Smale, Chair

In an attempt to understand DNA methylation in various contexts, we have examined chromatin modification at enhancers and CpG islands. At both DNA features, we find the binding of transcription factors is the major determinant of methylation status.

At the enhancer for the tissue-specific inflammatory gene *Il12b*, we attempted to isolate the DNA sequence necessary for the establishment of a low methylation window usually present in most cell types. We cloned *Il12b* enhancer deletions into a bacterial artificial chromosome that could recapitulate native chromatin when stably transfected into murine ES cells, but were unable to remove the low methylation window without deleting the full enhancer sequence. The *Il12b* enhancer is uniquely methylated in embryonic stem cells compared to all other cell types; it has higher methylation than usual and responds to certain changes in growth conditions. DNA

methylation increases globally during stem cell differentiation, but DNA methylation at the *Il12b* enhancer remains constant in successfully differentating cells. Finally, we take advantage of variable *Il12b* enhancer methylation in embryonic stem cells to demonstrate that, following differentiation to a macrophage fate, moderate enhancer methylation does not prevent *Il12b* expression.

To understand what factors influence the well studied low DNA methylation, histone 3 lysine 4 trimethylation (H3K4me3), and low nucleosome occupancy at CpG islands, we cloned CpG rich DNA into bacterial artificial chromosomes which were stably transfected into ES cells. Analysis of the integrated BACs revealed that CpG island features are each controlled through separate mechanisms. We determined several properties of CpG island features based on experimental deletions and fusions of a small CpG island and the 601 positioning sequence. Protection from DNA methylation at CpG islands can occur either by binding of a specific transcription factor, or by a size threshold mechanism in murine ES cells. H3K4me3 marking requires low DNA methylation, but unmethylated CpGs are not sufficient to recruit high levels. Nucleosome density is influenced by transcription factor binding and sequence positioning determinants, but is unaffected by low DNA methylation and moderate H3K4me3 levels.

We expanded our analysis of CpG islands to include all CpG rich regions in the human genome, which were computationally determined based on our own criteria. Using available chromatin datasets, we assayed the effect of nucleotide content on CpG island features. We found that CpG density and island size correlated with high levels of CpG island features. However, by far the strongest determinant of CpG island features was association with a promoter. Promoter CpG rich regions were strongly biased to accumulate high levels of all CpG island features, which could not be explained by nucleotide content. Instead, we showed that

promoter CpG islands have much higher transcription factor binding then other CpG islands in the genome, and high binding is correlated with lower DNA methylation and higher H3K4me3. Finally, we observed a difference in the DNA methylation at human and mouse CpG islands. High CpG density mouse CpG islands are much more susceptible to demethylation.

This multifaceted study elucidates many previously undefined relationships between transcription factors and chromatin properties. These findings will be beneficial to describing the complex mechanisms that drive regulation of cell fate and gene expression.

The dissertation of Justin Bryon Langerman is approved.

_____

Kenneth Dorshkind

_____

Michael F. Carey

_____

Stephen T. Smale, Committee Chair

University of California, Los Angeles

2014

Dedication

I would like to dedicate this research to each of my parents, for their inspiration, support, and encouragment.

Table of Contents

List of Figures and Tables

Acknowledgements

I would like to acknowledge my family for instilling in me early on the importance of education and curiosity. Thank you Mom, Dad, and Dad for your support through the years.

I would like to thank my mentor, Stephen Smale, for helping me grow as a scientist. Thanks for helping me to appreciate the rigor required in our field, for guiding me through it and for supporting my research interests.

I would like to thank my thesis committee, Douglas Black, Michael Carey, Ken Dorshkind, and Hanna Mikkola, for their attention and discussion regarding both my research project and education. I would also like to thank the rest of the Gene Regulation community at UCLA for their contribution to an invaluable education in my field of interest (and for all of the food!). I would also like to thank my previous mentors, Jau-Nian Chen and Jayoung Choi, for teaching me and starting me down the path.

I would like to thank many members of Smale lab past and present who have helped me and gave friendship freely. In particular, thank you to Jian Xu, Scott Pope, Kevin Doty, and Dev Bhatt for taking the time to build my technical expertise, and thank you to Miguel Edwards and Prabaht Purbey for the copious experimental discussion. Thanks to everyone in Smale lab for all the discussion, both germaine and irreverent, that got me through the years.

I would like to thank my wonderful girlfriend Priscilla for maintaining my sanity.

I would like to acknowledge the contributions of David Lopez to the work presented in Chapter 3. Thanks also to Michelle Lissner and Carly Schwartz for contributions to the project.

I would like to acknowledge the Cellular and Molecular Biology Training program for financial support during a portion of my graduate work. I would also like to thank the UCLA ACCESS program and Molecular Biology Institute for funding and support.

VITA

2006	Bachelors of Science University of California, Los Angeles Molecular, Cellular, and Developmental Biology
2004	Undergraduate Research Student Research Position, MCDB University of California, Los Angeles Dr. John Merriam
2005-2006	Undergraduate Research Student Research Position, MCDB University of California, Los Angeles Dr. Jau-Nian Chen
2006-2007	Laboratory Assistant Molecular, Cellular, and Developmental Biology University of California, Los Angeles Dr. Jau-Nian Chen
2007-2008	ACCESS Program University of California, Los Angeles
2008-2011	Ruth L. Kirstein National Research Service Award, Cellular and Molecular Training Grant University of California, Los Angeles
2009-2010	Teaching Assistant Microbiology, Immunology, and Molecular Genetics University of California, Los Angeles
2008-Present	Ph.D Candidate, Dr. Stephen T Smale Laboratory Molecular Biology Institute University of California, Los Angeles

Publications and Presentations

Establishment and Maintenance of Competence in the *Il12b* Enhancer During Pluripotency and Development
Poster Presentation, Seaborg Symposium, UCLA, 2010
Poster Presentation, UCLA MBI Retreat, 2012
Poster Presentation, Immunology LA, Skirball Institute, 2012

Determinants of the Fundamental Properties of CpG Islands in Pluripotent Cells,
Poster Presentation, 5th NIGMS Workshop on Pluripotent Stem Cell Research, NIH Bethesda, Maryland 2014

Chapter 1

Introduction

Regulation of DNA Methylation During Development and at CpG Islands

Chromatin and Gene Regulation

Epigenetics was coined to refer to non-genetic effects on phenotype but is now taken to mean either mitotically heritable non-DNA changes or more generally non-sequence based modifications that alter gene transcription (Bird, 2007; Peschansky and Wahlestedt, 2014; Waddington, 1942). This last category of epigenetics often refers to changes that target chromatin; the superstructure of DNA and proteins that organize and compact the genetic material. Accordingly, the study of epigenetic control of gene expression has been focused on modification to histones and DNA (Bernstein et al., 2006; Bird, 2002; Hajkova et al., 2008).

Deoxyribonucleic Acid

The basic structure of DNA is composed of nucleic acids or nucleotides, made up of four nitrogenous bases, the purines adenine and guanine and the pyrimidines cytosine and thymine, bound to a monosacchride sugar and a phosphate. The DNA polymer is formed by covalent bonds between the phosphate group and the sugar molecule of the next nucleotide. In all living organisms, two DNA polymers form a double helix structure, based on the hydrogen bonding of the DNA bases in adenine-thymine and cytosine-guanine pairs at the inside of the helix and twisting of the sugar-phosphate backbone on the outside (Watson and Crick, 1953). The seqeunce of the nucleobases encodes the biological information that makes up genetics.

Histone Modification

In eukaryotes, DNA is compacted by assembly into nucleosomes, accomplished by wrapping 147bp of DNA around the histone octamer. The histone octamer is composed of two of each of the core histone proteins H2A, H2B, H3 and H4 (Luger et al., 1997). Histones can be

post-translationally modified at both their DNA contacting core regions and at particular residues in the unstructured N-terminal tails, which protrude from the nucleosome and can readily be accessed (Li et al., 2007; Rogakou et al., 1998) Enzymatic activity on histones tails, such as acetylation on lysines which blocks the charge on the side group, can faciliate decondensation and promote accessibility (Reinke and Hörz, 2003). Transcriptional activation of some inducible genes requires decompaction of histones via acetylation by the acetyltransferase Gcn5 (Kuo et al., 1998). Histone modifications can also serve as recognition sites for binding or blocking of chromatin scanning proteins (Fischle et al., 2005).. These alterations, which include covalent addition of methyl, acetyl, ubiquitin, and phosphate groups to the histone tail amino acids, have defined roles in the cell depending on the residue modified and are collectively referred to as the "histone code" (Strahl and Allis, 2000).

Several important histone modifications have been well studied. Histone 3 Lysine 4 tri-methylation (H3K4me3) is generally associated with transcribing promoters (Santos-Rosa et al., 2002). Methyl groups may be added to Lysine 4 progressively as active promoters are also high for H3K4 mono-methylation and di-methylation. In yeast, it has been demonstrated that the histone methyltransferase Set1 is recruited by the initiating RNA Polymerase II complex and deposites H3K4me3 (Ng et al., 2003). The H3K4me3 modification also corresponds to active promoters in mammals, but several methyltransferases are involved; Set1, the MLL family, the ALL-1 complex, and Set7/9 (Lee and Skalnik, 2005; Nakamura et al., 2002; Wilson et al., 2002; Yokoyama et al., 2004). Mammalian H3K4 methylatransferases are often associated with large complexes containing general transcription factors, remodeling complexes, and histone acetylases/deacetylases.

Another critical modification is histone 3 lysine 27 tri-methylation (H3K27me3), a silencing modification first described as the enzymatic product of the Polycomb Complex 2 in Drosophilia (Czermin et al., 2002). This activity of this complex is crucial to silence developmental loci (Schwartz et al., 2006), and is dependent on the methyltransferase Enhancer of Zeste (Ebert et al., 2004). The polycomb complexes also exist in humans, where two homologues of Enhancer of Zeste, EZH1 and EZH2, both have H3K27 methyltransferase activity (Kuzmichev et al., 2002; Shen et al., 2008). H3K27me3 is bound by PRC1, which has numerous repressive enzymatic activities that can silence the local chromatin (Cao et al., 2002).

The same histone residue can have multiple modifications, for example the well studied histone 3 lysine 9 (H3K9) residue. H3K9 can be acetylated in active contexts or tri methylated, which strongly represses the surrounding chromatin and restricts access to control elements (Ayyanathan et al., 2003; Osipovich et al., 2004). H3K9 is an example of how modifications can counterbalance each other. Another example is when H3K4me3 and H3K27me3 are present at the same loci; the result is a bivalent state influenced by both the activating and silencing properties of the two marks (Bernstein et al., 2006). With dozens of known modifications and hundreds of potential sites, histones are a major contributor to the complex regulation of DNA access and transcription.

Modification of DNA by Methylation

The DNA molecule itself can also be covalently modified to alter the chromatin environment by transfer of a methyl group from donor substrate S-adenosylmethionine directly onto a nitrogenous base. This simple modification is found in most life on Earth, in both prokaryotes and eukaryotes (Marinus and Morris, 1973). In prokaryotes, methyltransferases like

Dam and Dcm modify the N^6 position of adenine or the C^5 position of cytosine respectively, as part of what is thought to be a steric based method of self recognition and protein-DNA contact control (Casadesús and Low, 2006). DNA methylation in higher eukaryotes is restricted to modification of the C^5 position of cytosine forming 5-methylcytosine, but is conserved across most animal kingdoms (Su et al., 2011). One class of DNA sequence is refractory to DNA methylation; large clusters of CpG dinucleotides are often unmethylated in mammals (Bird, 1985). These regions are termed 'CpG islands', and are an important regulatory feature of mammalian genomes that will be discussed in greater detail in later sections.

DNA Methylation in Mammals

In eukaryotes, 5-methylcytosine occurs most often in the context of cytosine-guanine dinucleotides (CpGs). Other contexts can be methylated, most notably in plants, where cytosines in a cytosine-non guanine sequence context can be targeted for methylation at nearly equal frequency to CpG (Hsieh et al., 2009). Recently it has been discovered that human and mouse DNA is also modified at these non-CpG contexts in certain cell types (Lister et al., 2009), but this event occurs rarely in comparison to CpG methylation and its function is not well understood. In animal genomes, CpG methylation is very common, as most CpGs are 60-80% methylated in most cell types. In mammals, this high level of modification is maintained by the constitutive activity of DNA methyltransferases (DNMTs), a family of five proteins in humans. DNMT1 is the maintenance methyltransferase, which maintains CpG methylation in a variety of tissues (Li et al., 1992). DMNT1's function as a maintenance methyltransferase is based on its strong preference for hemimethylated DNA over unmethylated DNA, allowing it to propagate methylation signals from parent DNA strand to daughter DNA strand (Fatemi et al., 2001).

DMNT3A and DMNT3B are considered de novo methyltransferases, and are especially active during embryogenesis where loss leads to embryonic lethality or severe birth defects (Okano et al., 1999). They are likely the mediators of non-CpG methylation in embryonic stem cells (Ramsahoye et al., 2000), and they also have different catalytic activites which lead DNMT3A and 3B to methylate different genomic targets (Gowher and Jeltsch, 2002). Two other homologous methyltransferases, DNMT2 and DNMT3L, play less direct roles in global DNA methylation. DNMT2 is highly conserved but has extremely weak activity *in vivo* and does not appear to be necessary for development (Liu et al., 2003). DNMT3L has no catalytic domain, but has been shown to interact with DNMT3a and 3b to methylate imprinting sites (Hata et al., 2002).

DNA Methylation and Nucleosomes

Another factor in DNA methylation may be higher order assembly into nucleosomes, but the relationship is currently ambiguous. Computational studies predict nucleosomes would require higher free energy to assemble on methylated DNA (Portella et al., 2013) and in vitro studies suggest that DNMTs prefer linker DNA and have difficulty methylating bases in contact with core histones or bound by the heterochromatin H1 linker histone (Takeshima et al., 2008), but in vivo studies in plants and humans found that nucleosomal DNA had higher methylation than linker DNA (Chodavarapu et al., 2010). While it is still inconclusive, a recent study looking at diverse eukaryotic species found that in algae DNA methylation is anticorrelated with nucleosome positioning (Huff and Zilberman, 2014), which suggests that the nucleosome-DNA methylation relationship is not necessarily predicated on universal histone-DNA interactions and may be due to species specific adaptations.

Mediating the DNA Methylation Signal

DNA methylation is commonly associated with heritable and stable repressive chromatin, and is sufficient to shut down gene expression (Doyes and Bird, 1991). This effect is achieved through two primary methods; steric hinderance and methyl-C binding proteins. Steric hinderance by the methyl group can interfere with the DNA binding of transcription factors. For some DNA binding proteins such as AP-1, methylation is sufficient to block binding and prevent transactivation (Comb and Goodman, 1990). However, the repressive chromatin environment usually seen at regions of high DNA methylation mostly results from the recognition of 5-methylcytosine by repressor proteins. This class of proteins that specifically binds methylated DNA and in mammals includes the closely related MBDs and MeCP2, which all have a common Methyl-C Binding Domain motif (Hendrich and Bird, 1998). The specificity of the MBD motif for methylated DNA is confered by interaction of 5 amino acids with the 5C methyl group (Ohki et al., 2001). MBD containing proteins are crucial for correctly mediating the silencing effect of DNA methylation; for instance mutation of MeCP2 is known to be responsible for the neurodevelopmental disorder Rett syndrome (Van den Veyver and Zoghbi, 2001).

Functional effects of the 5-methylcytosine signal can be created by recruitment of corepressor complexes. MBD2 associates with the NuRD complex which can repress transcription by nucleosome remodelling and histone deacetylation (Feng and Zhang, 2001). MeCP2 has a transcriptional repressor domain which can recruit Sin3a, an HDAC containing complex that represses transcription in a manner dependent on histone de-acetylase activity (Nan et al., 1998). MBD1 also has a motif capable of repressing in vitro transcription at methylated promoters (Fujita et al., 1999). MBD3 does not directly bind DNA but is part of the NuRD

complex. The repressive action of these complexes can also have positive feedback on DNA methylation; the heterochromatin protein HP-1 recruits DMNT1 (Smallwood et al., 2007).

DNA Methylation and Gene Silencing in Mammals

DNA methylation at promoters has long been associated with silencing of the associated genes (Boyes and Bird, 1992; Kass et al., 1997). Because of the repressive chromatin formation triggered by DNA methylation, it is often deposited to repress or maintain silencing of specific genomic targets. However DNA methylation does not act alone in gene silencing; it has been shown that it cannot be deposited at strong promoters (Boyes and Bird, 1992) or regions bound by activators (Macleod et al., 1994). This may be due, in part, to the antagonism between histone modifications that promote transcription and the DNA methylation machinery. It has been shown that Mll, an H3K4 methyltransferase, can antagonize DNA methylation where it binds (Erfurth et al., 2008). Similarly, it has been shown that DNMT3L, which helps to recruit DNMT3a for de novo methylation, can only bind histone tails depleted of H3K4 methylation (Ooi et al., 2007). Indeed, genome wide H3K4 methylation and DNA methylation are anticorrelated (Weber et al., 2007).

Gene silencing often occurs during development and differentiation, in which certain cell type related proteins have to be repressed. A notable example is repression of Oct-4, one of the master regulators of embryonic stem cell fate. Once differentiation is triggered, DNA sequences within the Oct-4 promoter are bound by a specific repressor family called the COUP-TFs, which block transcription (Schoorlemmer et al., 1994). In human neuronal precursors, Oct-4 transcription drops within four days of retinoic acid induced differentiation, but substantial DNA

methylation at the Oct-4 locus only begins to appear twelve days later (Deb-Rinker et al., 2005). Instead of precipitating repressive chromatin, DNA methylation generally occurs after other repressive proteins have targeted the locus. Further study of the link between repression of transcription and DNA methylation at the Oct4 locus found that the H3K9 methyltransferase G9a was capable of binding DNMT3a/b, and was required upstream of de novo methylation (Epsztejn-Litman et al., 2008). Interestingly, the catalytic activity of G9a was not required for DNMT3a recruitment, suggesting a direct interaction is sufficient. Other components may help to mediate this recruitment, like the ATPase LSH, which cooperates with the G9a upstream of DNA methylation at certain sites (Myant et al., 2011). H3K9me3 and DNA methylation stabilize silencing so that direct repressor binding is no longer required. This model was empirically demonstrated in experiments where pulse expression of a single repressor was sufficient to trigger stable and heritable H3K9me3 and DNA methylation deposition (Ayyanathan et al., 2003).

Female X Inactivation and DNA Methylation

One of the most remarkable silencing events in mammalian development is inactivation of the X chromosome in females, a complex process requiring silencing of an entire chromosome for dosage compensation. X inactivation actually occurs in two waves; a non random early imprinting of the paternal chromosome and a random inactivation during post-blastocyst growth that will result in a chimeric adult (Augui et al., 2011). The first inactivation occurs just after fertilization, where the oocyte protects the maternal X and silences "foreign" X chromosomes (Tada et al., 2000). Paternal silencing persists in the extra embryonic tissue but is quickly reversed in the embryo before subsequent inactivation. Random inactivation is dependent on the

X inactivation center, a region of the X chromosome that contains the Xist and Tsix non-coding RNAs and is necessary and sufficient to initiate chromosome silencing (Brown et al., 1991; Rastan, 1983). Transcripts of the highly stable Xist must necessarily coat the entire inactive X chromosome in cis (Clemson et al., 1996; Penny et al., 1996). This leads to recruitment of H3K27 trimethylation and the incorporation of histone variant H2A1, resulting in condensation of the inactive X into a very dense heterochromatin structure known as a macrochromatin body (Plath et al., 2003; Rasmussen et al., 2001). DNA methylation occurs at the inactive X as a late step to insure long term silencing; even normally resistant CpG islands are methylated (Gendrel et al., 2012; Norris et al., 1991). This late stage DNA methylation seems to be primarily a means of stabilizing the chromosomal silencing in somatic cells rather than initiating repression (Beard et al., 1995). Accordingly, knock out of DNA methyltransferase activity had no effect on X inactivation in differentiating ES cells (Sado et al., 2004). Across a range of targets, DNA methylation acts to stabilize heterochromatin at regions where silencing has already been initiated.

RNA directed DNA Methylation

While most CpGs in the vertebrate genome are moderately methylated and change methylation state gradually, there exists a large dynamic range of variation for many genomic features. The exception to normal gene silencing, described above, is when specific regions of the genome are targeted for DNA methylation. One of the best described mechanisms of targeted DNA methylation is the control of transposons and repetetive regions. Transposons are ancient selfish DNA elements found in all eukaryotes that can replicate and insert themselves into new positions in the genome (Wicker et al., 2007). Suppressing replication of these elements is

crucial for the cell, as they can become a significant genomic burden; in some plant species transposons comprise more than half of the genome (Phillips, 1998). Unregulated transposon replication can even result in sterility (Lin and Spradling, 1997).

Both plants and animals use specific RNA based targeting of DNA methylation to combat the spread of transposons (Kim and Zilberman, 2014; Siomi et al., 2011). In plants, RNA Dependent RNA Polymerase 6 creates double stranded RNA from active transposon transcripts which complex with Argonautes to recruit Domain Rearranged Methyltransferase to genomic transposon locations. The resulting DNA methylation is then stably maintained by RNA Polymerase IV directed DNA methylation, discussed below (Nuthikattu et al., 2013; Wassenegger et al., 1994).

PIWI directed DNA Methylation

Animals have a similar transposon silencing system, first discovered in Drosophila, that is based on a subset of the RNAi machinery specific to transposons. The complex involved is the highly conserved Argonaute-like PIWI proteins and piwi-RNAs (Cox et al., 1998; Vagin et al., 2006). The PIWI proteins are Auburgine, Piwi, and Ago3 in flies, MILI and MIWI in mice, and PIWILs in human. The highest PIWI expression occurs in the germ line (Kuramochi-Miyagawa et al., 2001). The piRNA are small RNA of 24-30nt which are transcribed from transposon rich clusters in a primary wave and then undergo replication for a secondary wave resulting in an amplified library of piRNAs sense and antisense to transposon transcripts (Aravin et al., 2008). The piRNA 2’ end contains an O-methyl group which is recognized by the PIWI argonauts but prevents binding to the rest of the Argonaute clade (Tian et al., 2011). The PIWI-RNA complex localizes to the nucleus where is shuts down transposon transcription and results in DNA

methylation of repeat elements across the genome (Kuramochi-Miyagawa et al., 2008), although the exact mechanism remains unclear. There is evidence that PIWI directed DNA methylation can also exist outside the germline, such as in the central nervous system where it plays a role in shut down of CREB2 in response to serotonin (Rajasethupathy et al., 2012).

Plant Specific RNA Directed DNA Methylation

Plants also have RNA-targeted DNA methylation that silences regions other than insertion elements. In Arabadopsis, methylation and silencing of the flowering gene FWA requires the RNAi machinery, RNA dependent RNA polymerase, and RNA Polymerase IV and V, for a process termed RNA directed DNA methylation (Chan et al., 2004; Onodera et al., 2005; Zheng et al., 2007). RNA Polymerase IV transcribes a long ssRNA which is amplified by RDR2 and processed into siRNA and associated with the Argonaute. This targeting unit forms a complex which recruits the plant DNA methyltransferase families Drm and Cmt in addition to HDACs and chromatin remodelers to the original PolIV transcription site (Cao et al., 2003; Kanno et al., 2004; Matzke et al., 2009). RNA Polymerase V acts to stabilize regions silenced by this pathway (Mosher et al., 2008). Notably, RNA directed DNA methylation of non-transposons does not seem to occur in animals. For instance, DNA methylation at centromeric and alpha satellite regions seems to be promoted in response to H3K9 tri methylation instead of a direct targeting mechanism (Peters et al., 2001).

Epigenetic Reprogramming of DNA Methylation During Development

DNA methylation is stable in somatic tissues, but undergoes complete removal and re-establishment during mammalian development (Reik et al., 2001). These events, precipitated by

formation of the germ line and fertilization, erase the targeted deposition of DNA methylation and reprogram the epigenetic landscape of the cell. Each fertilized embryo actually undergoes two waves of demethylation (Monk et al., 1987). The first event occurs in primordial germ cells (PGCs), by E14 in mice, during which all genomic methylation is actively removed and totipotency is restored. After complete demethylation, the maternal and the paternal PGC will soon accrew DNA methylation again at many sites (Kafri et al., 1992).

Reprogramming in Primordial Germ Cells

It is during this time in the primordial germ cells that DNA methylation must be established at imprinted genes (Tucker et al., 1996). Imprinted genes are a special class of mono-allelic transcripts that are always specifically expressed from the maternal or paternal chromosome (Feil et al., 1994). Regulation of imprinted loci can be quite complex, involving long range interactions and barrier proteins (Szabó et al., 2004), but the hallmark of imprinted loci is stably inherited DNA methylation at the control locus. Imprinting gene control is necessary for correct regulation of embryonic development; disruption can lead to disorders such as Angelmann's and Praeder Willi syndromes in humans (Cattanach et al., 1992). DNMT3a is responsible for the de novo methylation at imprinting sites, while DNMT3b is relegated to methylation at satellite and centromeric DNA (Kaneda et al., 2004). Once the correct methylation patterns are established, the PGCs will develop into mature germ cells in the adult organism and undergo meiosis before fertilization.

Reprogramming After Fertilization

Upon fertilization of an oocyte with a sperm, both the paternal and maternal pronuclei undergo another wave of demethylation, with sudden paternal demethylation and gradual maternal demethylation preceeding nuclei fusion (Mayer et al., 2000). This phase does not erase all methylation, such as the imprinting established during PGC development (Olek and Walter, 1997). Loss of methylation from the paternal pronucleus takes place in under four hours, coupled to sperm decondensation and the loss of protamine. Upon nuclei fusion but before cleavage, the zygote has very low total genomic 5-methylcytosine and de novo methylation begins (Santos et al., 2002). With the exception of the trophoblast extra-embryonic tissue, total methylation grows during cleavage, reaching levels comparable to somatic tissues by the blastocyst stage. In humans, trophoblast tissue will acquire methylation over development but remains about 15% lower globally until birth (Schroeder et al., 2013). After these waves of epigenetic reprogramming the embryo will correctly develop.

Mechanism of Reprogramming Related DNA Demethylation

Recently, the mechanism of active demethylation during these reprogramming events has become clearer. It had long been speculated that passive loss of methylation by dilution during DNA replication, when coupled with inhibition of DMNT1, could be responsible for demethylation but several studies found that active demethylation occurred without DNA replication (Hajkova et al., 2008; Mayer et al., 2000). Instead, a compelling alternative was discovered when

5-hydroxymethylcytosine was found in mammalian DNA, a version of 5-methylcytosine modified with a hydroxyl group by the Tet (ten eleven translocation) family of dioxygenase

proteins (Tahiliani et al., 2009). The existence of this mark quickly led to description of an active pathway for demethylation; 5-methylcytosine is modified to 5-hydroxymethylcytosine by the Tet family, and then modified again to 5-carboxylcytosine. This new modification is quickly recognized by Thymine DNA-Glycosylase and undergoes base exision repair, replacing the nucleotide with an unmethylated cytosine (He et al., 2011). This process has been shown to be responsible for the active demethylation that occurs during PGC formation (Hackett et al., 2013), finally explaining how at least some active demethylation occurs during epigenetic reprogramming.

5-hydroxymethylcytosine formation is an important mechanism outside of reprogramming as well, where it plays a role in tasks like maintenance of low methylated regions that allows gene expression (Ito et al., 2010).

Pluripotency and Embryonic Stem Cells

One of the hallmarks of epigenetic control is its role in the maintenance of stem populations. Stem cells are capable of division and self-renewal, but can also differentiate into the specific tissues of the body. Embryonic stem cells(ESCs), isolated from the inner cell mass of a pre-implantation embryo during the blastocyst stage, are the only culturable cell population capable of differentiating into any tissue of the body (Evans and Kaufman, 1981; Thomson et al., 1998). This property is referred to as pluripotency. ESCs can contribute to a chimeric mouse when injected into an embryo (Bradley et al., 1984) and can be maintained indefinitely in culture (Xie et al., 2010). ESCs are an excellent system for research because they are ammenable to genetic manipulation and can be induced to differentiate *in vivo* to a number of somatic cell types.

Pluripotency Transcription Factor Network

Principal for maintaining pluripotency in ESCs is a network of DNA binding transcription factors. The most important of these factors is the gene product of Pou5f1, the octamer binding protein Oct-4 (Schöler and Dressler, 1990). Loss of Oct4 is embryonic lethal and causes spontaneous differentiation of isolated inner-cell mass cells from knockout embryos (Nichols et al., 1998). Supporting Oct-4 in the pluripotency network are a number of proteins including Nanog, Sox2, Esrrb, Tbx3, Tcl1, and Dppa4 (Ivanova et al., 2006). In vertebrates, Oct-4, Sox2, and Nanog tend to bind together at the early zygotic genes (Leichsenring et al., 2013), but also bind their own promoters to maintain expression in an autoregulatory circuit (Boyer et al., 2005). When differentiation is triggered, silencing of pluripotency network proteins like Oct-4 and Nanog by repressors is critical for exit from pluripotency (Cole et al., 2008; Schoorlemmer et al., 1994).

In the Nobel prize winning experiment from Yamanaka's group, it was demostrated that introduction of a group of these core pluripotency factors into somatic cells was sufficient to induce dedifferentiation to a pluripotent state (Takahashi and Yamanaka, 2006; Takahashi et al., 2007). These induced pluripotent stem (iPS) recapitulated ES cell state and function, including the ability to contribute to a chimeric mouse (Okita et al., 2007). Originally, the iPS cells were created by viral transduction with Oct-4, Nanog, Sox2, and c-Myc. Later it was shown that different mixtures of the pluripotency network proteins could induce reprogramming (Nakagawa et al. 2008; Schwarz et al. 2014; Buganim et al. 2012). Most of these mixtures act through or activate Oct-4, while the other network factors are interchangeable (Radzisheuskaya and Silva, 2014). Indeed, iPS cells can be derived solely by overexpression of Oct-4, although only in cell

types that express some of the pluripotency network factors already (Kim et al., 2009; Tsai et al., 2011).

Chromatin in Pluripotent Cells

Pluripotent cells are known to have a unique chromatin environment compared to somatic cells. In general their chromatin is more accessable; histology reveals they have fewer heterochromatin foci than somatic cells (Aoto et al., 2006; Meshorer et al., 2006). ES cells also have less repressive histone modifications compared to somatic cells, with 3-4 fold less total H3K9me3 and H3K27me3 (Hawkins et al., 2010). Repressive histone marks in ES cells are more likely to be mitigated by activatory histone marks (Bernstein et al., 2006); these bivalent domains are much less common in other tissues (Rugg-Gunn et al., 2010). Perhaps as a consequence of a more accessable chromatin environment, ES cells also have a detectable increase in global transcription from both genic and non-genic regions (Efroni et al., 2008).

Degrees of Pluripotency

Pluripotency of the embryo exists on a gradient as it develops, evidenced by the isolation and characterizeation of murine post-implantation embryonic cells called epiblast stem cells (Epi-SCs) which have numerous differences from ESCs (Brons et al., 2007; Tesar et al., 2007). EpiSCs can form teratomas but cannot contribute to a chimeric mouse and lack some markers of ESCs like alkaline phosphatase staining. Additionally the pluripotency network differs between EpiSCs and ESCs; Oct4 binds different targets, the EpiSC transcriptome contains more genes related to germ layer differentiation, and the pluripotency maintenance network uses different

proteins. Interestingly, EpiSCs could only be derived once human ESC culturing conditions were tried.

EpiSCs help explain differences between mouse ESCs and human ESCs. Human ESCs have different culturing requirements, due to a necessity for SMAD2/3 activation of the activin/nodal pathway to maintain pluripotency as opposed to the murine LIF/STAT3 pathway (James et al., 2005). Human ESCs also have different targets for their pluripotency factors; only about 10% of targets are also bound in mouse ESCs (Loh et al., 2006). Additionally, human ESCs have very active DNA repair machinery which makes them much more prone to apoptosis than murine ESCs (Qin et al., 2007). This variation can be explained by placing murine ESCs, human ESCs, and EpiSCs on a pluripotency continuum from "naïve", to "primed." Further evidence of the plasticity of pluripotency came from the discovery that addition of GSK and MEK inhibitors (2i media) could increase the resistance of murine ES cells to differentiation (Wray et al., 2011). Briefly, this effect is caused by stimulation of Wnt self-renewal signaling by GSK inhibition, and by abrogation of a side effect of LIF usage by its removal; LIF also stimulates Erk signaling in addition to STAT3 which is a pro-differentiation signal (Kunath et al., 2007; Sato et al., 2004). Supporting the validity of a continuum, two groups showed that human ESCs could be stably reprogrammed to a naïve state where they closely resembled mouse ESCs (Gafni et al., 2013; Ware et al., 2014). This was accomplished by treatment of prederived human ES lines with a 16 factors targeting the pluripotency network in addition to 2i media, or by treatment with histone deacetylase inhibitors and 2i media. These experiments support the conclusions that differentiation is a step wise process of increasing specification and loss of plasticity. The ability of a histone deacetylase inhibitor to alter human ESCs is yet more proof that epigenetics plays an important role in cell fate specification.

Distribution of DNA Methylation in ES cells and During Development

Analysis of DNA methylation in the sequencing era is based on bisulfite conversion of the DNA base cytosine into cytosine-sulphonate, which can be reduced to uracil by ammonia. Bisulfite attacks the 5-carbon in the cytosine ring, meaning 5-methylcytosine is protected from conversion. Sequencing of bisulfite treated DNA distinguishes methylation state in this manner, bypassing the need for special restriction sites or copious amounts of radiation (Frommer et al., 1992). With the addition of deep sequencing technology, bisulfite sequencing provides base pair resolution of the methylation status of nearly every CpG in the genome (Laurent et al., 2010; Lister et al., 2009; Stadler et al., 2011; Vincent et al., 2013; Ziller et al., 2013).

Genome Wide Methylomes

Surprisingly, the genomic DNA methylation average in ES cells is 15% higher than IMR90 fibroblasts (Lister et al., 2009). Cytosine methylation in a non-CpG context was identified and is largely ES cell specific. Across all genes, DNA methylation is generally very low at promoters, and then higher than genomic average throughout the gene body. Ranking ES genes by expression revealed that the lowest promoter methylation correlated with the highest gene expression (Laurent et al., 2010). Though it may be counterintuitive, higher gene body methylation is also correlated with the highest expressed genes.

DNA Methylation and Transcription Factor Binding Sites

Transcription factor binding sites and cell type specific enhancers have a low DNA methylation footprint (Lister et al., 2009; Ziller et al., 2013). Conversely, identification of small

regions with low DNA methylation genome wide can identify transcription factor bound DNA (Burger et al., 2013; Feldmann et al., 2013; Stadler et al., 2011). The majority of low methylated regions overlapped with DNase hypersensitivity sites, another method for measuring DNA accessability and inferring transcription factor binding. Interogation of the DNA bound to the transcription factor CTCF by ChIP-bisulfite sequencing found that the methylation status of bound DNA is always low (Feldmann et al., 2013). Furthermore, 5-hydroxymethyl cytosine was found at these transcription factor targets, suggesting active demethylation was occuring.

These small, low methylated regions often change methylation status during differentiation, and the underlying motifs at these regions are generally cell type specific (Ziller et al., 2013). Many of these sites also contain the enhancer related histone modification H3K27 acetylation. Examination of tissue specific enhancers provides evidence that DNA methylation can indicate epigenetic regulation well before gene transcription. The tissue specific genes *Albumin, Prtca,* and *Il-12b* all have enhancer regions with low methylation windows in ES cells (Xu et al., 2007). Upon differentiation into relevant cell types these low methylation windows may expand, but during pluripotency and differentiation the tissue specific enhancers remain bound by factors that protect from DNA methylation. Additionally, pre-methylation of plasmids containing the tissue specific enhancers revealed that only pluripotent ES cells were capable of establishing a low methylated region (Xu et al., 2009). Theoretically, DNA binding activity at enhancers during pluripotency could protect a region from default DNA methylation and lineage restriction. Once a cell developes into a determinant lineage, its newly expressed factors can bind the pre-accessable chromatin (Samstein et al., 2012). Most of the putative enhancer regions caught be genome wide studies seem to be lineage restricted, based on DNA methylation changes, but some of the tissue specific genes in (Xu et al., 2007) are protected across many cell

types. DNA methylation may provide a key to understanding how different classes of enhancers are regulated across differentiation.

CpG Islands

One feature of mammalian genomes that is especially refractory to DNA methylation is the accumulation of CpGs at unusually high density. CpG dinucleotides are depleted in the genome four to five fold below expected by random distribution, probably due to the mutagenic effects of spontaneous deamination of 5-methyl C to uracil (Shen et al., 1992, 1994). A deaminated nucleotide will be targeted by mismatch repair where it can be repaired back to a cytosine or to thymine depending on the template strand chosen. Conversion to TpG is irreversible, eliminating the CpG and the possibility for DNA methylation at that site. However, large stretches of CpG rich DNA, referred to as "CpG islands," have escaped genomic depletion (Bird, 1985; Gardiner-Garden and Frommer, 1987). CpG islands have a unique chromatin environment; they were originally discovered due to their consistent low DNA methylation which led to frequent cleavage by the methylation sensitive HPI restriction enzyme. In addition to unique chromatin properties, CpG islands can be found at 70% of coding gene promoters and nearly all house-keeping genes, where they presumably contribute to transcriptional regulation (Davuluri et al., 2001). In humans and mice, nearly 10% of genes share a large CpG island at their promoter with another gene (Engström et al., 2006). Looking across the genome in several species, CpG density above background correlates with species complexity; the greatest enrichment of CpG islands is in mammals, with minimal enrichment in inverterbrates, and no evidence for CpG islands in *E. Coli* (Irizarry et al., 2009). Interestingly, some plants have regions of high CpG density which are often near genes, suggesting convergent evolution on regulatory

CpG islands (Ashikawa, 2001). Regulation of CpG islands is important for control of cell fate (Fouse et al., 2008), and misregulation is a frequent event in cancer (Hinoue and Weisenberger, 2012).

Defining CpG Islands

Although CpG islands have been discovered for three decades and search guidelines have been evolving, there is still no consensus on definition. Many proposed definitions have used a sequence based approach that takes advantage of the unusual CpG density at CpG islands, based on the original Gardiner-Garden equation to define the observed over expected CpG number:

(Number of CpGs * Island Size) / (Number of G's and C's)

The putative region is considered a CpG island if its value is greater than 0.6 (Gardiner-Garden & Frommer 1987). Since then, this criteria has been adjusted to find the most promoter associated CpG islands and minimize repeat elements and methylated regions (Davuluri et al., 2001; Kent et al., 2002; Saxonov et al., 2006; Takai and Jones, 2002). One issue with this approach is it requires GC content and size cutoffs to function, which creates false negatives. The cutoffs used by UCSC Genome Browser, the standard for CpG island definition, are currently a 300bp size, 0.55 Observed/Expected and a 55% GC content cutoff.

With the increasing availability of DNA methylation and histone modification datasets, some searches have been adjusted to define and classify CpG islands based on favorable chromatin criteria (Bock et al., 2007; Fan et al., 2008). These studies still require a baseline input sequence, and therefore often start from the above computational criteria. When researching promoters several groups use low, medium, and high CpG content designations; this avoids using cut-offs but requires pre-selection of sequence for study (Fenouil et al., 2012; Landolin et

al., 2010). For non-biased calling of CpG islands, the most far reaching technique is to base the annotation on regions that are distinct from neighboring DNA, for instance by scanning for CpG dincleotide density with computational algorythms (Irizarry et al., 2009). Another approach is to take advantage of CpG chromatin; one group has adapted a protein that binds unmethylated CpGs to create an unmethylated CpG affinity column (Illingworth et al., 2010). This enriches specifically for large amounts of unmethylated CpGs and therefore CpG island DNA which can be sequenced. Using this approach, it was shown that half of the unmethylated CpG islands in the human and mouse genomes are not near genic locations and these orphan CpG islands are more sensitive to differentiation initiated DNA methylation.

Chromatin Features of CpG Islands

No precise definition exists for a CpG island, but they are frequently associated with several chromatin features. Although 60-80% of CpGs in the mammalian genome are methylated (Lister et al., 2009), the CpGs within CpG islands often remain completely resistant to modification (Suzuki and Bird, 2008). CpG islands are also strongly correlated with the activatory histone modification H3K4 trimethylation regardless of location (Mikkelsen et al., 2007). In accordance with the relationship between H3K4me3 and the transcriptional machinery, study of large unmethylated CpG islands has found that they are also often associated with the transcription machinery, even at intergenic sites (Illingworth et al., 2010).

Bivalent domains were first described at CpG islands promoters, where the usual H3K4me3 mark is accompanied by deposition of the repressive histone mark H3K27me3 (Azuara et al., 2006; Bernstein et al., 2006). Although most research concentrates on bivalent domains in ES cells, they have also been discovered at promoters in hematopoietic progenitors

and in mature T cells (Cui et al., 2009; Roh et al., 2006). Bivalent CpG islands in ES cells are often associated with genes important for development, where initiating RNA polymerase II remains paused due to activity of Polycomb repressive complex 1 (Brookes et al., 2012). Additionally, some non-developmental bivalent promoters have low levels of transcription that are increased by knockdown of Polycomb repressive complex 1. This balancing act between positive and negative regulators of transcription is thought to "poise" chromatin for correct expression once development begins. After differentiation, these bivalent domains will resolve, losing H3K27me3 in the correct cell types. Bivalent CpG islands can also resolve into repression for lineage restricted of genes not appropriate to the new cell type, in which case the CpG island retains H3K27me3 and undergoes DNA methylation (Mohn et al., 2008).

CpG Islands and Nucleosome Occupancy

CpG islands are often depleted of nucleosomes when they are located near the promoters of active genes in mammals (Fenouil et al., 2012). This is perhaps unsurprising, as it has been known for some time that depletion and decompaction of nucleosomes promote transcription (Han and Grunstein, 1988; Kuo et al., 1998), but the question remains whether the nucleosome depletion is due to DNA sequence or transcriptional influences. In vitro experiments have shown that DNA sequences can contain positioning determinants; selection for tight bindings sequences resulted in a 150bp sequence named 601 that has the strongest known affinity for nucleosomes (Lowary and Widom, 1998). The histone binding strength of 601 is conferred by rigid guanine and cytosine tracts periodically interrupted by flexible thymine-adenine dinucleotides (Fernandez and Anderson, 2007; Vasudevan et al., 2010). In yeast almost all promoters have a nucleosome free region, but the size of the depletion varies with the strength of gene activity (Weiner et al.,

2010). Interestingly, minimal digestion with the micrococcal nuclease used during mapping revealed that these nucleosome free regions actuality still contained nucleosomes with fragile positioning, which are easily evicted by transcription initiation (Xi et al., 2011). Yeast does not have CpG islands, but a simple model where high guanine-cytosine content (GC%) can provide protection from poly-A tracks explains a majority of the *in vivo* positioning data (Tillo and Hughes, 2009).

In humans, nucleosome positions are very reponsive to transcription factor binding and transcription initiation (Fu et al., 2008; Schones et al., 2008). Indeed, some inducible genes like the tissue specific Interleukin gene *Il12b* can undergo sudden nucleosome remodelling at their promoters and enhancers in response to external stimuli (Ramirez-Carrozzi et al., 2006; Zhou et al., 2007). However, CpG islands are correlated with reduction of nucleosome occupancy and with decreased dependence on nucleosome remodeling machinery in mammals (Ramirez-Carrozzi et al., 2009). As CpG islands necessarily have high GC content, mammalian nucleosome positioning is not behaving like yeast positioning. One study, looking in depth at nucleosome mapping in mouse cells, found that increased CpG content and GC content lowered the average nucleosome occupancy at promoters ranked by RNA Polymerase II binding (Fenouil et al., 2012). Inhibition of Pol II with α-amanitin resulted in an increase of nucleosome density at the borders of GC rich regions but not over the core sequence, suggesting both DNA sequence and transcriptional events influence CpG island nucleosome occupancy. In human cells, nucleosome mapping and then clustering of occupancy patterns at most promoters revealed extreme heterogeneity in the borders and shapes of nucleosome free regions (Kundaje et al., 2012). The same study found that high GC content could predict the nucleosome footprint around transcription factor binding sites for the proteins CTCF and SP1, but only if regions with

similar asymmetric footprints were clustered allowing small GC content similarities to become apparent. Deciphering the logic driving nucleosome depletion at CpG islands and why sequence determinants seem to change in different contexts will most likely require further classification and reduction.

Mechanisms for Acquisition of CpG Island Features

CpG islands seem to acquire their chromatin features autonomously. One study showed that transcription factor binding sites within CpG islands were necessary and sufficient to establish their low DNA methylation status (Lienert et al., 2011). After insertion into the β-globin locus in ES cells, pieces of the Nanog promoter CpG island were able to maintain low DNA methylation, and were correctly methylated upon differentiation to neural progenitors. Mutation of known transcription factor binding sites in small inserts abrogated protection from methylation. The idea of protective DNA binding is supported by observations that widely bound transcription factors like CTCF and REST are responsible for many small pockets of low DNA methylation distal to promoters in the genome (Stadler et al., 2011). One study inserted CpG rich *E. Coli* sequence, which should be deficient for conserved mammalian transcription factor binding sites, into the β-globin locus. They found that 70% of the *E. Coli* inserts acquired DNA methylation in mammalian cells, with only the most CpG dense inserts escaping heavy DNA methylation and acquiring H3K4me3 (Lienert et al., 2011). Another study also using CpG rich *E. Coli* insertions into ES cells found that CpG rich DNA acquired the activatory H3K4me3 modification regardless of origin (Mendenhall et al., 2010). However, the E Coli DNA also acquired the repressive H3K27me3 modification. In their model, histone modification does not seem to be driven by specific sites as any CpG rich DNA is capable of acquiring H3K4me3 and

H3K27me3, but for some CpG islands Polycomb deposition of H3K27me3 can be blocked by the presence of activatory transcription factor binding. Both groups focused on separate chromatin modifications, but their models generally agree that CpG rich DNA can intrinsically acquire repressive chromatin modifications unless protected by activatory transcription factor binding sites.

Another explanation for the enrichment of H3K4me3 at CpG islands arose from observation that the CXXC motif containing protein Cfp1 is bound to most CpG islands (Thomson et al., 2010). The CXXC motif of Cfp1 can only bind unmethylated CpG DNA, but also contains homology to H3K4 methyltransferase proteins (Voo et al., 2000). Knockdown of Cfp1 reduces H3K4me3 at promoter CpG islands, suggesting that it is a likely mediator of this CpG island feature (Thomson et al., 2010). In this study they also introduced exogenous CpG rich DNA into a genomic locus. The Puro-EGFP cassette was inserted into the 3' UTR of the Nanog gene, where is acquired low DNA methylation and both Cfp1 binding and high H3K4me3.

These studies have improved the understanding of CpG island features a great deal, but there is still tension in the current models. To date no study has considered each feature at CpG islands; most groups only look at one or two features. We therefore do not know how the binding site mutations in (Lienert et al., 2011) affect H3K4me3, or if nucleosome occupancy is reduced at *E. Coli* DNA insertions in ES cells. The thresholds of CpG density and size necessary to trigger these features are also poorly understood. However, now is a better time than ever to be researching the subtle and complex inputs that control CpG island evolution and epigenetics. Genome wide datasets and consortiums like the ENCODE project provide an incredible armorment of support to conventional experiments (Bernstein et al., 2012). Future studies will

establish definitive understanding of the nucleotide and environmental requirements underlying acquisition of CpG island features.

References

Aoto, T., Saitoh, N., Ichimura, T., Niwa, H., and Nakao, M. (2006). Nuclear and chromatin reorganization in the MHC-Oct3/4 locus at developmental phases of embryonic stem cell differentiation. Dev. Biol. *298*, 354–367.

Aravin, A. a, Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol. Cell *31*, 785–799.

Ashikawa, I. (2001). Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. Plant J. *26*, 617–625.

Augui, S., Nora, E.P., and Heard, E. (2011). Regulation of X-chromosome inactivation by the X-inactivation centre. Nat. Rev. Genet. *12*, 429–442.

Ayyanathan, K., Lechner, M.S., Bell, P., Maul, G.G., Schultz, D.C., Yamada, Y., Tanaka, K., Torigoe, K., and Rauscher, F.J. (2003). Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. Genes Dev. *17*, 1855–1869.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nat. Cell Biol. *8*, 532–538.

Beard, C., Li, E., and Jaenisch, R. (1995). Loss of methylation activates Xist in somatic but not in embryonic cells. Genes Dev. *9*, 2325–2334.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bird, A. (1985). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Bird, A. (2007). Perceptions of epigenetics. Nature *447*, 396–398.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG Island Mapping by Epigenome Prediction. PLoS Comput. Biol. *3*, e110.

Boyer, L. a, Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Boyes, J., and Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. EMBO J. *1*, 327–333.

Bradley, A., Evans, M., Kaufman, M., and Robertson, E. (1984). Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. Nature.

Brons, I.G.M., Smithers, L.E., Trotter, M.W.B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R. a, et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191–195.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., et al. (2012). Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell *10*, 157–170.

Brown, C., Ballabio, A., and Rupert, J. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature.

Buganim, Y., Faddah, D. a, Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. *41*, e155.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039–1043.

Cao, X., Aufsatz, W., Zilberman, D., Mette, M.F., Huang, M.S., Matzke, M., and Jacobsen, S.E. (2003). Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation. Curr. Biol. *13*, 2212–2217.

Casadesús, J., and Low, D. (2006). Epigenetic gene regulation in the bacterial world. Microbiol. Mol. Biol. Rev. *70*, 830–856.

Cattanach, B., Barr, J., and Evans, E. (1992). A candidate mouse model for Prader–Willi syndrome which shows an absence of Snrpn expression. Nat. ….

Chan, S.W., Zilberman, D., Xie, Z., Johansen, L.K., Carrington, J.C., and Jacobsen, S.E. (2004). B REVIA RNA Silencing Genes Control de. Nature *303*, 2004.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y. V, Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J. a, Kuo, F., Kim, J., Cokus, S.J., et al. (2010). Relationship between nucleosome positioning and DNA methylation. Nature *466*, 388–392.

Clemson, C.M., McNeil, J. a, Willard, H.F., and Lawrence, J.B. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J. Cell Biol. *132*, 259–275.

Cole, M.F., Johnstone, S.E., Newman, J.J., Kagey, M.H., and Young, R. a (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes Dev. *22*, 746–755.

Comb, M., and Goodman, H.M. (1990). CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic Acids Res. *18*, 3975–3982.

Cox, D.N., Chao, a., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev. *12*, 3715–3727.

Cui, K., Zang, C., Roh, T.-Y., Schones, D.E., Childs, R.W., Peng, W., and Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell *4*, 80–93.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell *111*, 185–196.

Davuluri, R. V, Grosse, I., and Zhang, M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. *29*, 412–417.

Deb-Rinker, P., Ly, D., Jezierski, A., Sikorska, M., and Walker, P.R. (2005). Sequential DNA methylation of the Nanog and Oct-4 upstream regions in human NT2 cells during neuronal differentiation. J. Biol. Chem. *280*, 6257–6260.

Doyes, J., and Bird, A. (1991). DNA Methylation via a Methyl-CpG Inhibits Transcription Binding Protein Indirectly. *64*, 1123–1134.

Ebert, A., Schotta, G., Lein, S., Kubicek, S., Krauss, V., Jenuwein, T., and Reuter, G. (2004). Su(var) genes regulate the balance between euchromatin and heterochromatin in Drosophila. Genes Dev. *18*, 2973–2983.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., Le Grice, S., McKay, R.D.G., Buetow, K.H., et al. (2008). Global transcription in pluripotent embryonic stem cells. Cell Stem Cell *2*, 437–447.

Engström, P.G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S.L., Yang, L., et al. (2006). Complex Loci in human and mouse genomes. PLoS Genet. *2*, e47.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., Cedar, H., et al. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nat. Struct. Mol. Biol. *15*, 1176–1183.

Erfurth, F.E., Popovic, R., Grembecka, J., Cierpicki, T., Theisler, C., Xia, Z.-B., Stuart, T., Diaz, M.O., Bushweller, J.H., and Zeleznik-Le, N.J. (2008). MLL protects CpG clusters from methylation within the Hoxa9 gene, maintaining transcript expression. Proc. Natl. Acad. Sci. U. S. A. *105*, 7517–7522.

Evans, M., and Kaufman, M. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature.

Fan, S., Zhang, M.Q., and Zhang, X. (2008). Histone methylation marks play important roles in predicting the methylation status of CpG islands. Biochem. Biophys. Res. Commun. *374*, 559–564.

Fatemi, M., Hermann, a, Pradhan, S., and Jeltsch, a (2001). The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA. J. Mol. Biol. *309*, 1189–1199.

Feil, R., Walter, J., Allen, N.D., and Reik, W. (1994). Developmental control of allelic methylation in the imprinted mouse Igf2 and H19 genes. Development *120*, 2933–2943.

Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. *9*, e1003994.

Feng, Q., and Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deactylating methylated nucleosomes. Genes Dev. *15*, 1031–1040.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., et al. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. *22*, 2399–2408.

Fernandez, A.G., and Anderson, J.N. (2007). Nucleosome positioning determinants. J. Mol. Biol. *371*, 649–668.

Fischle, W., Tseng, B., and Dormann, H. (2005). Regulation of HP1–chromatin binding by histone H3 methylation and phosphorylation. Nature *438*.

Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell Stem Cell *2*, 160–169.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U. S. A. *89*, 1827–1831.

Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. *4*, e1000138.

Fujita, N., Takebayashi, S., Kudo, S., Chiba, T., Saya, H., and Okumura, K. (1999). Methylation-Mediated Transcriptional Silencing in Euchromatin by Methyl-CpG Binding Protein MBD1 Isoforms Methylation-Mediated Transcriptional Silencing in Euchromatin by Methyl-CpG Binding Protein MBD1 Isoforms.

Gafni, O., Weinberger, L., Mansour, A.A., Manor, Y.S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. Nature *504*, 282–286.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Gendrel, A.-V., Apedaile, A., Coker, H., Termanis, A., Zvetkova, I., Godwin, J., Tang, Y.A., Huntley, D., Montana, G., Taylor, S., et al. (2012). Smchd1-dependent and -independent pathways determine developmental dynamics of CpG island methylation on the inactive X chromosome. Dev. Cell *23*, 265–279.

Gowher, H., and Jeltsch, A. (2002). Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases. J. Biol. Chem. *277*, 20409–20414.

Hackett, J. a, Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T. a, and Surani, M.A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science *339*, 448–452.

Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U.C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M.A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. Nature *452*, 877–881.

Han, M., and Grunstein, M. (1988). Nucleosome loss activates yeast downstream promoters in vivo. Cell *55*, 1137–1145.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development *129*, 1983–1993.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell *6*, 479–491.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333*, 1303–1307.

Hendrich, B., and Bird, A. (1998). Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *18*.

Hinoue, T., and Weisenberger, D. (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome … 271–282.

Hsieh, T.-F., Ibarra, C. a, Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R.L., and Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. Science *324*, 1451–1454.

Huff, J.T., and Zilberman, D. (2014). Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. Cell *156*, 1286–1297.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet. *6*, e1001134.

Irizarry, R., Wu, H., and Feinberg, A. (2009). A species-generalized probabilistic model-based definition of CpG islands. Mamm. Genome *20*, 674–680.

Ito, S., D'Alessio, A.C., Taranova, O. V, Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature *442*, 533–538.

James, D., Levine, A.J., Besser, D., and Hemmati-Brivanlou, A. (2005). TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. Development *132*, 1273–1282.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, a (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev. *6*, 705–714.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.

Kanno, T., Mette, M., Kreil, D., and Aufsatz, W. (2004). Involvement of putative SNF2 chromatin remodeling protein DRD1 in RNA-directed DNA methylation. Curr. Biol. *14*, 801–805.

Kass, S.U., Landsberger, N., and Wolffe, a P. (1997). DNA methylation directs a time-dependent repression of transcription initiation. Curr. Biol. *7*, 157–165.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, a. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Kim, M.Y., and Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. Trends Plant Sci. 1–7.

Kim, J.B., Sebastiano, V., Wu, G., Araúzo-Bravo, M.J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrich, M., van den Boom, D., et al. (2009). Oct4-induced pluripotency in adult neural stem cells. Cell *136*, 411–419.

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development *134*, 2895–2902.

Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C.L., Raha, D., Winters, E.E., Johnson, S.M., Snyder, M., Batzoglou, S., and Sidow, A. (2012). Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res. *22*, 1735–1747.

Kuo, M., Zhou, J., Jamberk, P., Churchill, M., and Allis, C. (1998). Histone acetyltransferase activity of yeast Gcn5p is required for the activation of target genes in vivo. Genes Dev. 627–639.

Kuramochi-Miyagawa, S., Kimura, T., Yomogida, K., Kuroiwa, a, Tadokoro, Y., Fujita, Y., Sato, M., Matsuda, Y., and Nakano, T. (2001). Two mouse piwi-related genes: miwi and mili. Mech. Dev. *108*, 121–133.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. *22*, 908–917.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev. *16*, 2893–2905.

Landolin, J.M., Johnson, D.S., Trinklein, N.D., Aldred, S.F., Medina, C., Shulha, H., Weng, Z., and Myers, R.M. (2010). Sequence features that drive human promoter function and tissue specificity. Genome Res. 890–898.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Lee, J.-H., and Skalnik, D.G. (2005). 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.

Leichsenring, M., Maes, J., Mössner, R., Driever, W., and Onichtchouk, D. (2013). Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science *341*, 1005–1009.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097.

Lin, H., and Spradling, a C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development *124*, 2463–2476.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Liu, K., Wang, Y.F., Cantemir, C., and Muller, M.T. (2003). Endogenous Assays of DNA Methyltransferases : Evidence for Differential Activities of DNMT1 , DNMT2 , and DNMT3 in Mammalian Cells In Vivo Endogenous Assays of DNA Methyltransferases : Evidence for Differential Activities of DNMT1 , DNMT2 , and DNMT3 in M.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet. *38*, 431–440.

Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. *276*, 19–42.

Luger, K., Mäder, A., and Richmond, R. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *7*, 251–260.

Macleod, D., Charlton, J., Mullins, J., and Bird, a P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282–2292.

Marinus, M.G., and Morris, N.R. (1973). Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. J. Bacteriol. *114*, 1143–1150.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., and Matzke, A.J.M. (2009). RNA-mediated chromatin-based silencing in plants. Curr. Opin. Cell Biol. *21*, 367–376.

Mayer, W., Niveleau, a, Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501–502.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P.J., Brown, D.T., and Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Dev. Cell *10*, 105–116.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell *30*, 755–766.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371–382.

Mosher, R. a, Schwach, F., Studholme, D., and Baulcombe, D.C. (2008). PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. Proc. Natl. Acad. Sci. U. S. A. *105*, 3145–3150.

Myant, K., Termanis, A., Sundaram, A.Y.M., Boe, T., Li, C., Merusi, C., Burrage, J., de Las Heras, J.I., and Stancheva, I. (2011). LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. Genome Res. *21*, 83–94.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat. Biotechnol. *26*, 101–106.

Nakamura, T., Mori, T., Tada, S., Krajewski, W., Rozovskaia, T., Wassell, R., Dubois, G., Mazo, A., Croce, C.M., and Canaani, E. (2002). ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. Mol. Cell *10*, 1119–1128.

Nan, X., Ng, H.H., Johnson, C. a, Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, a (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Ng, H.H., Robert, F., Young, R. a, and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol. Cell *11*, 709–719.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H., and Smith, a (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell *95*, 379–391.

Norris, D.P., Brockdorff, N., and Rastan, S. (1991). Methylation status of CpG-rich islands on active and inactive mouse X chromosomes. Mamm. Genome *1*, 78–83.

Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N., and Slotkin, R.K. (2013). The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. Plant Physiol. *162*, 116–131.

Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M., and Shirakawa, M. (2001). Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. Cell *105*, 487–497.

Okano, M., Bell, D.W., Haber, D. a, and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature *448*, 313–317.

Olek, A., and Walter, J. (1997). The pre-implantation ontogeny of the H19 methylation imprint. Nat. Genet.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell *120*, 613–622.

Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717.

Osipovich, O., Milley, R., Meade, A., Tachibana, M., Shinkai, Y., Krangel, M.S., and Oltz, E.M. (2004). Targeted inhibition of V(D)J recombination by a histone methyltransferase. Nat. Immunol. *5*, 309–316.

Penny, G., Kay, G., and Sheardown, S. (1996). Requirement for Xist in X chromosome inactivation. Nature.

Peschansky, V.J., and Wahlestedt, C. (2014). Non-coding RNAs as direct and indirect modulators of epigenetic regulation. Epigenetics *9*, 3–12.

Peters, a H., O'Carroll, D., Scherthan, H., Mechtler, K., Sauer, S., Schöfer, C., Weipoltshammer, K., Pagani, M., Lachner, M., Kohlmaier, a, et al. (2001). Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. Cell *107*, 323–337.

Phillips, R.L. (1998). Grass genomes. Proc. Natl. Acad. Sci. U. S. A. *95*, 1975–1978.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K. a, Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science (80-.). *300*, 131–135.

Portella, G., Battistini, F., and Orozco, M. (2013). Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. PLoS Comput. Biol. *9*, e1003354.

Qin, H., Yu, T., Qing, T., Liu, Y., Zhao, Y., Cai, J., Li, J., Song, Z., Qu, X., Zhou, P., et al. (2007). Regulation of apoptosis and differentiation by p53 in human embryonic stem cells. J. Biol. Chem. *282*, 5842–5852.

Radzisheuskaya, A., and Silva, J.C.R. (2014). Do all roads lead to Oct4? The emerging concepts of induced pluripotency. Trends Cell Biol. *24*, 275–284.

Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., Tuschl, T., and Kandel, E.R. (2012). A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell *149*, 693–707.

Ramirez-Carrozzi, V.R., Nazarian, A. a, Li, C.C., Gore, S.L., Sridharan, R., Imbalzano, A.N., and Smale, S.T. (2006). Selective and antagonistic functions of SWI/SNF and Mi-2beta nucleosome remodeling complexes during an inflammatory response. Genes Dev. *20*, 282–296.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, a P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. U. S. A. *97*, 5237–5242.

Rasmussen, T.P., Wutz, a P., Pehrson, J.R., and Jaenisch, R.R. (2001). Expression of Xist RNA is sufficient to initiate macrochromatin body formation. Chromosoma *110*, 411–420.

Rastan, S. (1983). Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre. J. Embryol. Exp. Morphol. *78*, 1–22.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science *293*, 1089–1093.

Reinke, H., and Hörz, W. (2003). Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. Mol. Cell *11*, 1599–1607.

Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S., and Bonner, W.M. (1998). DNA Double-stranded Breaks Induce DNA Double-stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139. J. Biol. Chem. *273*, 5858–5868.

Roh, T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proc. Natl. Acad. Sci. U. S. A. *103*, 15782–15787.

Rugg-Gunn, P.J., Cox, B.J., Ralston, A., and Rossant, J. (2010). Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. Proc. Natl. Acad. Sci. U. S. A. *107*, 10783–10790.

Sado, T., Okano, M., Li, E., and Sasaki, H. (2004). De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. Development *131*, 975–982.

Samstein, R.M., Arvey, A., Josefowicz, S.Z., Peng, X., Reynolds, A., Sandstrom, R., Neph, S., Sabo, P., Kim, J.M., Liao, W., et al. (2012). Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell *151*, 153–166.

Santos, F., Hendrich, B., Reik, W., and Dean, W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev. Biol. *241*, 172–182.

Santos-Rosa, H., Schneider, R., and Bannister, A. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407–411.

Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. (2004). Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. Nat. Med. *10*, 55–63.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. *103*, 1412–1417.

Schöler, H., and Dressler, G. (1990). Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. EMBO J. *9*, 2185–2195.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Schoorlemmer, J., Puijenbroek, M., Eijnden, M., Jonk, L., Pals, C., and Kruijer, W. (1994). Characterization of a negative retinoic acid response element in the murine Oct4 promoter. Mol. Cell. Biol. *14*.

Schroeder, D.I., Blair, J.D., Lott, P., Yu, H.O.K., Hong, D., Crary, F., Ashwood, P., Walker, C., Korf, I., Robinson, W.P., et al. (2013). The human placenta methylome. Proc. Natl. Acad. Sci. U. S. A. *110*, 6037–6042.

Schwartz, Y.B., Kahn, T.G., Nix, D. a, Li, X.-Y., Bourgon, R., Biggin, M., and Pirrotta, V. (2006). Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat. Genet. *38*, 700–705.

Schwarz, B. a, Bar-Nur, O., Silva, J.C.R., and Hochedlinger, K. (2014). Nanog is dispensable for the generation of induced pluripotent stem cells. Curr. Biol. *24*, 347–350.

Shen, J., III, W.R., and Jones, P. (1992). High frequency mutagenesis by a DNA methyltransferase. Cell *71*, 1073–1080.

Shen, J.C., Rideout, W.M., and Jones, P. a (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. *22*, 972–976.

Shen, X., Liu, Y., Hsu, Y.-J., Fujiwara, Y., Kim, J., Mao, X., Yuan, G.-C., and Orkin, S.H. (2008). EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. Mol. Cell *32*, 491–502.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A. a (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nat. Rev. Mol. Cell Biol. *12*, 246–258.

Smallwood, A., Estève, P., Pradhan, S., and Carey, M. (2007). Functional cooperation between HP1 and DNMT1 mediates gene silencing. Genes Dev. *21*, 1169–1178.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Strahl, B., and Allis, C. (2000). The language of covalent histone modifications. Nature *403*, 41–45.

Su, Z., Han, L., and Zhao, Z. (2011). Conservation and divergence of DNA methylation in eukaryotes: New insights from single base-resolution DNA methylomes. Epigenetics *6*, 134–140.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. *9*, 465–476.

Szabó, P.E., Tang, S.E., Silva, F.J., Tsark, W.M.K., and Mann, J.R. (2004). Role of CTCF Binding Sites in the Igf2 / H19 Imprinting Control Region Role of CTCF Binding Sites in the Igf2 / H19 Imprinting Control Region. Mol. Cell. Biol. *24*.

Tada, T., Obata, Y., Tada, M., Goto, Y., Nakatsuji, N., Tan, S., Kono, T., and Takagi, N. (2000). Imprint switching for non-random X-chromosome inactivation during mouse oocyte growth. Development *127*, 3101–3105.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W. a, Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science (80-.). *324*, 930–935.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Takai, D., and Jones, P. a (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. U. S. A. *99*, 3740–3745.

Takeshima, H., Suetake, I., and Tajima, S. (2008). Mouse Dnmt3a preferentially methylates linker DNA and is inhibited by histone H1. J. Mol. Biol. *383*, 810–821.

Tesar, P.J., Chenoweth, J.G., Brook, F. a, Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

Thomson, J. a., Itskovitz-Eldor, J., Shapiro, S., Waknitz, M., Swiergiel, J., Marshall, V., and Jones, J. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. Science (80-.). *282*, 1145–1147.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Tian, Y., Simanshu, D.K., Ma, J.-B., and Patel, D.J. (2011). Structural basis for piRNA 2'-O-methylated 3'-end recognition by Piwi PAZ (Piwi/Argonaute/Zwille) domains. Proc. Natl. Acad. Sci. U. S. A. *108*, 903–910.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Tsai, S., Bouwman, B., Ang, Y., and Kim, S. (2011). Single transcription factor reprogramming of hair follicle dermal papilla cells to induced pluripotent stem cells. Stem … 964–971.

Tucker, K.L., Beard, C., Dausmann, J., Jackson-Grusby, L., Laird, P.W., Lei, H., Li, E., and Jaenisch, R. (1996). Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. Genes Dev. *10*, 1008–1020.

Vagin, V. V, Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science *313*, 320–324.

Vasudevan, D., Chua, E.Y.D., and Davey, C. a (2010). Crystal structures of nucleosome core particles containing the "601" strong positioning sequence. J. Mol. Biol. *403*, 1–10.

Van den Veyver, I.B., and Zoghbi, H.Y. (2001). Mutations in the gene encoding methyl-CpG-binding protein 2 cause Rett syndrome. Brain Dev. *23 Suppl 1*, S147–51.

Vincent, J.J., Huang, Y., Chen, P.-Y., Feng, S., Calvopiña, J.H., Nee, K., Lee, S. a, Le, T., Yoon, A.J., Faull, K., et al. (2013). Stage-specific roles for tet1 and tet2 in DNA demethylation in primordial germ cells. Cell Stem Cell *12*, 470–478.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., and Skalnik, D.G. (2000). Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase , Human Trithorax , and Methyl-CpG Binding Domain Protein 1 Cloning of a Mammalian Transcriptional Activator That Binds.

Waddington, C. (1942). Canalization of development and the inheritance of acquired characters. Nature.

Ware, C.B., Nelson, A.M., Mecham, B., Hesson, J., Zhou, W., Jonlin, E.C., Jimenez-Caliani, A.J., Deng, X., Cavanaugh, C., Cook, S., et al. (2014). Derivation of naive human embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *111*, 4484–4489.

Wassenegger, M., Heimes, S., Riedel, L., and Sänger, H.L. (1994). RNA-directed de novo methylation of genomic sequences in plants. Cell *76*, 567–576.

Watson, J., and Crick, F. (1953). Molecular structure of nucleic acids. Nature.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

Weiner, A., Hughes, A., Yassour, M., Rando, O.J., and Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res. *20*, 90–100.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

Wilson, J.R., Jing, C., Walker, P. a, Martin, S.R., Howell, S. a, Blackburn, G.M., Gamblin, S.J., and Xiao, B. (2002). Crystal structure and functional analysis of the histone methyltransferase SET7/9. Cell *111*, 105–115.

Wray, J., Kalkan, T., Gomez-Lopez, S., Eckardt, D., Cook, A., Kemler, R., and Smith, A. (2011). Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. Nat. Cell Biol. *13*, 838–845.

Xi, Y., Yao, J., Chen, R., Li, W., and He, X. (2011). Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. Genome Res. *21*, 718–724.

Xie, X., Hiona, A., and Lee, A. (2010). Effects of long-term culture on human embryonic stem cell aging. Stem Cells Dev. *20*.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J. a, Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *104*, 12377–12382.

Xu, J., Watts, J. a, Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev. *23*, 2824–2838.

Yokoyama, A., Wanh, Z., Wysocka, J., Mrinmoy, S., Aufiero, D., Kitabayashi, I., Herr, W., and Cleary, M. (2004). Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression. Mol. Cell. Biol. *24*, 5639–5649.

Zheng, X., Zhu, J., Kapoor, A., and Zhu, J.-K. (2007). Role of Arabidopsis AGO6 in siRNA accumulation, DNA methylation and transcriptional gene silencing. EMBO J. *26*, 1691–1701.

Zhou, L., Nazarian, A. a, Xu, J., Tantin, D., Corcoran, L.M., and Smale, S.T. (2007). An inducible enhancer required for Il12b promoter activity in an insulated chromatin environment. Mol. Cell. Biol. *27*, 2698–2712.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D. a, Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.

Chapter 2

Establishment and Maintenance of Low DNA Methylation at the *Il12b* Enhancer

Abstract

At some genes, cell-type specific transcription factor binding is potentiated during pluripotency by pioneer factors which can mark bound enhancers with low methylation windows. We intensely studied one such region, the *Il12b* enhancer, to determine how the low methylation window is controlled. The low methylation at the *Il12b* enhancer is established cooperatively by nearly the entire enhancer sequence. We also found that the *Il12b* enhancer is uniquely regulated in embryonic stem cells, with variable methylation that can respond to growth conditions and cell state changes. Differentiation of ES cells demonstrates that methylation at the *Il12b* enhancer may remains constant in cells that successfully reach the new lineage, but does not correlate with the expression level of the *Il12b* gene.

Introduction

Transcription factor binding to DNA is an essential component of gene regulation and control of cell fate. The most famous transcription factors are those involved in the self regulatory pluripotency network; Oct-4, Sox2 and Nanog (Boyer et al., 2005; Loh et al., 2006). The activity of these transcription factors is so critical that their expression is sufficient to reprogram somatic cells to a pluripotent stage (Nakagawa et al., 2008; Takahashi and Yamanaka, 2006).

Essential to our understanding of transcription factors is the fact that their binding often perturbs local chromatin, creating barriers or disrupting nucleosomes (Burger et al., 2013; Felsenfeld et al., 1996). A simple event like cleavage of DNA by DNase I can therefore provide multitudes of information about the cellular environment including the ability to determine the cell's lineage (Thurman et al., 2012). Transcription factors can also induce modifications to the

local chromatin, creating areas of low DNA methylation (Stadler et al., 2011) or locally modifying histones (Heintzman et al., 2007).

Understanding the function of transcription binding can be a complex issue. Unraveling of the β-globin locus, a classic model of transcription factor binding functioning at enhancers at long range, slowly moved from discovery of DNase hypersensitivity sites to description of the locus control region and cell type specific enhancers (Grosveld et al., 1987; Li et al., 2002). Another sophisticated transcription factor binding event occurs at the enhancer for the tissue specific gene *Il12b*. In response to lipopolysacchride, a bacterial component, mature macrophages will acquire DNase hypersensitivity roughly 10kb upstream of the *Il12b* promoter. This regulatory enhancer region is essential for inducing high levels of the inflammatory gene transcript (Zhou et al., 2007). Before stimulation of mature macrophages, the *Il12b* enhancer is quietly associated with a nucleosome. Upon stimulation, a regulatory cascade induces remodeling at the enhancer in a matter of hours (Ramirez-Carrozzi et al., 2006).

Studies of DNA methylation during differentiation have proposed the idea that binding of enhancers by transcription factors is a binary event which happens upon lineage commitment (Stadler et al., 2011; Zhu et al., 2013). However, careful study of tissue specific enhancers does reveal binding activity before lineage commitment in some cases. The crucial regulatory T cell factor FoxP3 was found to bind sites that were made accessible before FoxP3 expression, including sites pre-bound by the homologue FoxO1 (Samstein et al., 2012). The muscle cell transcription factor MyoD binds to many myotube specific enhancers in predecessor myoblasts (Blum et al., 2012). For certain enhancers, the chromatin is bound and poised for recruitment of the transcription factors which will drive expression upon lineage commitment.

We have reported that the enhancers of the tissue specific genes *Il12b, Ptcra,* and *Albumin* contained unmethylated CpGs in embryonic stem(ES) cells (Xu et al., 2007). DNA methylation is high genome wide (Lister et al., 2009), so the low DNA methylation at the *Il12b* enhancer indicates transcription factor binding well before any transcriptional activity occurs. At the liver specific *Albumin* gene, an unmethylated CpG occurs in a crucial binding site for the transcription factor FoxA1. In ES cells, this site was found to be bound by the homologuous FoxD3. This is evidence that the poising of a tissue-specific enhancer may occur at the start of development, well before expected.

FoxD3's presense at the *Albumin* enhancer in ES cells suggested it had the characteristics of a pioneer factor: a protein that binds early in development to facilitate later binding by cell type specific factors. Indeed, knockdown of FoxD3 seems to lead to restriction of the *Albumin* enhancer as it loses the low methylation at the FoxA1 binding site (Xu et al., 2009). Transcription factors were also shown to be important in maintenance of the low methylation at the enhancer of the *Ptcra* gene. Dissection of the binding sites at the *Ptcra* enhancer showed that a number of positive and negative regulators of methylation cooperated to control the boundaries of the low methylation window in ES cells.

Although the factors which control the low methylation window at the *Il12b* enhancer are unknown, it was shown that establishment of the unmethylated CpGs must occur during pluripotency. Transfection of a pre-methylated plasmid containing the *Il12b* enhancer into a somatic cell line resulted in stable retention of high methylation. However, transfection into embryonic stem cells resulted in re-establishment of the low methylation window for both the *Il12b* enhancer and the *Ptcra* enhancer. The unique environment of pluripotent ES cells is critical for the establishment of a poised *Il12b* enhancer (Xu et al., 2009).

As the actual factors binding the *Il12b* enhancer in ES cells had not been described, we began a mutagenic assay to isolate the DNA sequences necessary for establishment and maintenance of the low methylation window. We found that most of the roughly 1kb *Il12b* enhancer region had some ability to cause demethylation locally. Additionally, we found that the *Il12b* enhancer methylation in ES cells could be variable, unlike in somatic cells. Overall our studies demonstrate the complexity of regulation occuring at the *Il12b* enhancer locus during development.

Results

The *Il12b* Enhancer Has Broad Demethylating Capability

To isolate the *Il12b* enhancer sequence necessary for establishment of the low DNA methylation window in ES cells, we cloned a series of plasmids with broad deletions of the full enhancer. Initial studies of the *Il12b* enhancer subdivided it into 5 fragments, A-E. The C fragment contains Oct and C/EBP binding sites crucial for transcription, while the D-E fragments contain the nearest CpGs and the low methylation window. We cloned enhancer fragments into an *Il12b* promoter plasmid and stably tranfected murine ES cells (Fig 2-1a). Without the presence of any *Il12b* enhancer DNA, the majority of clones spontaneously became heavily methylated (Fig 2-2a). However, introduction of the *Il12b* enhancer C fragment was sufficient to induce low levels of methylation at surrounding CpGs (Fig 2-1b). The C fragment triggered low methylation even when the plasmid used for transfection was pre-methylation *in vitro* with SssI. Similar low methylation was instigated by a strong constitutively active enhancer, the CpG rich hCMV enhancer sequence (Fig 2-2b).

Addition to the C enhancer of the DE fragments, which contains the 6 CpG *Il12b* low methylation window, seemed to slightly mitigate the activity of the C fragment (Fig 2-1c). Surprisingly, the DE fragment alone still had low methylation at the integrated plasmid *in vivo*, promoting low methylation in 3 out of 5 clones (Fig 2-1d). Although the C fragment alone results in the lowest methylation, the DE fragment also contains demethylation potential, suggestion that the *Il12b* enhancer is broadly bound by protective transcription factors. However, another possibility is that the DNA methylation levels at integrated plasmids are strongly affected by position effect variagation and by the nearly CpG island like character of the vector backbone. For this reason, studies were continued in bacterial artificial chromosomes (BACs).

Partial *Il12b* Enhancer Deletion in BACs Minimally Alters DNA Methylation

BACs are modifyable DNA that can be large enough to include a complete gene and its associated control regions (Heintz, 2001). Because of their large size, BACs can assemble into native chromatin and are buffered from position effect variagation in most cases. This provides an ideal environment for the study of the low DNA methylation window at the *Il12b* enhancer. Our initial approach was to make several deletions in the *Il12b* enhancer region overlapping the unmethylated CpGs, and at the previously characterized upstream C/EBP and Oct sites in the C fragment (Fig 2-3a). At the enhancer deletion *Il12b* BAC, which lacks the the C/EBP and Oct binding sites in addition to most of the DE fragment, we find the remaining adjacent CpGs have high methylation(Fig 2-3b). This suggests that the binding sites crucial for low methylation are within the *Il12b* CE enhancer fragment sequence. The C/EBP and Oct mutations did not seem to have a substantial effect on the *Il12b* enhancer window (Fig 2-3c,d). Although the methylation at some CpGs was higher, for both mutations the average methylation across the enhancer window

was 4-6% lower than at the endogenous *Il12b* enhancer. Complicating phenotypic characterization is the fact that the endogenous *Il12b* enhancer has a surprisingly variable methylation state in ES cells, although it still remains 10-30% below genomic background at its highest. One unambiguous observation was the methylation status of the CpG introduced by substitution mutation of the Oct binding site; it was 100% methylated in both clones tested despite being directly adjacent to the low methylation window at the *Il12b* enhancer.

Next we analyzed deletions within the *Il12b* enhancer window. We deleted an 87bp sequence which contains binding sites for Nfκb and the hematopoietic transcription factor Evi-1. This deletion does not remove CpGs. In the Evi-1 enhancer deletion BAC, we find that the low methylation at the *Il12b* enhancer remains largely unaffected in ES cells (Fig 2-4a). The average DNA methylation across four clones at the Evi-1 mutant BAC enhancer was 42% compared to 48% at the endogenous *Il12b* enhancer locus. The lack of effect suggested that larger deletions may be necessary. A 240bp deletion of the first half of the *Il12b* enhancer DE fragment removes two of six CpGs from the low methylation window. Bisulfite sequencing of the Half-DE deletion *Il12b* BAC reveals that the remaining CpGs remains largely unaffected, as we see only a slight increase in methylation at the BAC enhancer (Fig 2-4b). Direct comparison of CpGs covered by sequencing shows that the Half-DE deletion enhancer CpGs have 11% higher DNA methylation than the equivalent endogenous CpGs. Suprisingly, deletion of the reciprical 251bp second half of the DE enhancer fragment also has only a small effect on the remaining two CpGs (Fig 2-4c). The equivalent CpGs have 50% average DNA methylation in the Half-DE 2 deletion BAC compared to 43% at the endogenous *Il12b* enhancer. In each of the deletions tested, the endogenous *Il12b* enhancer had variably higher than expected DNA methyation in ES cells which complicated analysis.

Il12b Enhancer Methylation is Uniquely Increased in Pluripotent Cells

The DNA methylation variation at the *Il12b* enhancer in ES cells is particularly curious, as primary cells and somatic cell lines very clearly have stable and low DNA methylation at the same locus in many cell types (Fig 2-5a) (Xu et al., 2007). The endogenous *Il12b* enhancer in primary cells is generally less than 20% methylated at the first four CpGs in the DE fragment, but in ES cells we find methylation ranges from 20-80% with an average of approximately 45% for the same four CpGs. To determine whether the culture conditions could possibly be responsible, several facets of ES cell growth and culture media were tested. First, we tested the possibility that the mitotically inactivated mouse embryonic fibroblast feeder layer used to support ES cell growth was contaminating the bisulfite sequencing. Bisulfite sequencing of the *Il12b* enhancer in the the feeder independent ES line CCE and in R1 ES cells depleted of MEFs via replating demonstrates that the variable DNA methylation is still present (Fig 2-5b). Inactivated MEFs themselves display very low methylation at the *Il12b* enhancer, making them an unlikely source of variation. We tested several splitting conditions but were unable to find any particular growth protocol to restore low DNA methylation (Fig 2-5c). Altering oxidation conditions in the culture media, by increasing the amount of β-mercaptoethanol, also did not lower the *Il12b* enhancer methylation in ES cells (Fig 2-5d). High DNA methylation at the *Il12b* enhancer can be seen in genome wide bisulfite sequencing ES cell datasets, suggesting that an aberrant genetic event in our ES cells is not likely to be responsible for this phenomena (Fig 2-5e).

Re-evalution of Mutant BACs In ES Cells with Lower DNA Methylation

Although there was no growth condition which brought ES cell methylation at the *Il12b* enhancer in line with primary cells, we did find that changing the growth serum from undefined fetal bovine serum to defined knockout serum resulted in a stable 15-30% reduction in DNA methylation at the endogenous locus (Fig 2-6a). This allowed us to reassess the large deletion BACs which had shown subtle effects earlier. ES stable lines with the Half-DE and Half DE 2 Enhancer Deletion BACs were grown in knockout serum media and retested by bisulfite sequencing. The Half-DE enhancer deletion had moderately different methylation in this condition, with an average methylation 30% higher at the BAC enhancer than equivalent endogenous CpGs (Fig 2-6b). Bisulfite sequencing of Half-DE 2 enhancer deletion BAC lines in knockout serum media revealed a similar though smaller effect (Fig 2-6c). The average DNA methylation for shared CpGs was 35% at the BAC enhancer compared to 16% at the endogenous enhancer. These large deletions demonstrate an effect on the low methylation window, but the subtlety of both Half DE deletions suggest either functional redundancy in maintenance of the low methylation window, or contribution from an adjacent region.

The nearby region most likely to contribute is the C fragment of the *Il12b* enhancer, based on its demethylation activity in plasmids. More BACs were cloned with large deletions overlaping the C fragment, and then they were stably transfected into ES cells (Fig 2-7a). Both of the C-deletion BAC lines were grown in knockout serum media. Deletion of the C and D *Il12b* enhancer fragments has a large effect on the two remaining CpGs in the low methylation window (Fig 2-7b). Overall this effect strongly resembles the first BAC tested, the full enhancer deletion BAC. The second C deletion BAC combined C fragment deletion with the first Half DE deletion, a construct that keeps 4 enhancer CpGs (Fig 2-7c). The C-half DE enhancer deletion BAC has

slightly higher enhancer methylation than the endogenous locus, similar to the Half DE deletion alone. The deletion BAC had an average methylation of 37% compared to 23% at equivalent CpGs in the endogenous enhancer.

Moderate Methylation at the *Il12b* Enhancer Is Maintained Through Differentiation

To take advantage of the moderate methylation seen in ES cell lines in fetal bovine serum culture media, we decided to test the effect of variable *Il12b* enhancer methylation on *Il12b* expression. ES cells can be differentiated into a pure macrophage population capable of expressing immunity genes, including *Il12b* (Moore et al., 1998)(Pope et al., Unpublished Data). Bisulfite sequencing was used to monitor the methylation status of the *Il12b* enhancer window in wild type ES cell lines before and after differentiation. For this experiment we utilized the R1 ES line, the ROSA V6.5 ES line, and an induced pluripotent stem cell line (IPS). Each pluripotent cell line has a different methylation status at the *Il12b* enhancer (Fig 2-8a). ROSA ES cells were consistently high, with an average CpG methylation of 82% across the *Il12b* enhancer, while R1 cells were moderate at 53% and IPS cells were low at 27%. After differentiation to macrophages, methylation levels at the *Il12b* enhancer fell slightly for ROSA and IPS cells, but not R1 ES cells (Fig 2-8b). IPS cells, which started low, were the only cells post differentiation to achieve a low enhancer methylation window that was similar to that seen in primary cells. R1 and ROSA ES cells both had moderately methylated *Il12b* enhancers after differentiation, with approximately 50% average CpG methylation. ES derived macrophages were stimulated with lipopolysaccride to induce expression of *Il12b*, which was measured by qPCR (Fig 2-8c). All ES lines differentiated to macrophages expressed *Il12b* at appropriate levels, in addition to the inflammatory genes *Il-6* and *RANTES*. *Il12b* enhancer methylation state did not correlate with

expression of the *Il12b* gene. R1 macrophages expressed *Il12b* over 2 fold higher than the ROSA and IPS lines. Additionally R1 macrophages had uniquely high *Il-6* expression, while ROSA macrophages had uniquely low *RANTES* expression. We conclude that moderate methylation at the *Il12b* enhancer does not functionally prevent *Il12b* expression.

We next asked whether the *Il12b* enhancer methylation changed during differentiation to macrophages, as we have previously observed very high methylation in mid-differentiation embryoid bodies (Xu et al., 2007). To test this hypothesis, we attempted to compare successfully differentiating cells to stalled cells at three stages of the macrophage differentiation protocol. The first stage is at the third day of embryoid body formation, when a small population of $cKit^+$ cells which may become hematopoietic progenitors are first detectable. Bisulfite sequencing of these early progenitors shows no difference from total embryoid bodies at the *Il12b* enhancer window (Fig 2-8d). At day 6 of embryoid body formation, $Ckit^+$ $CD41^+$ cells make up large hematopoietic progenitor portion of the embryoid body population. In sorted cells the *Il12b* enhancer has an average CpG methylation of 41%, compared to 63% in total embryoid body cells. The final separation takes place at the early Macrophage I stage, after embryoid body disruption and treatment of the hematopoietic progenitors with Il-3. The proto-macrophages become suspension cells while the majority of the non-hematopoietic cells become adherent, and can easily be separated. Bisulfite sequencing comparing the suspension cells to the adherent cells reveals that the suspension cells have 13% lower methylation at the *Il12b* enhancer. While the differences are extremely slight, there is no evidence that the cells which successfully become macrophages change enhancer methylation status over the course of differentiation. Instead, it seems that cells which cannot successfully differentiate are likely to have increased methylation at the *Il12b* enhancer.

Discussion

In an attempt to isolate the crucial sequence for establishment of the low methylation window at the *Il12b* enhancer, we have discovered more complexity at the locus during pluripotency and differentiation. Plasmid studies showed that most of the *Il12b* enhancer sequence can cause demethylation locally and at surrounding CpGs, in ES cells. The C fragment of the *Il12b* enhancer, which contains no CpGs, seems to have the strongest demethylation ability. However, the C-E and D-E enhancer plasmids yielded extremely similar results, both with higher local DNA methylation than the C fragment alone. This suggests repressive factors may bind in the DE sequence to mitigate the strength of the C fragment binding sites. There is precedence for this at the *Ptcra* enhancer, where repressive proteins like Myb bind to limit the enhancer window size (Xu et al., 2009). A major caveat of the plasmid system is the possibility for unnatural chromatin formation. The vector backbone is extremely CpG rich, which may trigger chromatin changes (Mendenhall et al., 2010). Plasmids are also susceptable to changes resulting from their insertion sites (Allshire et al., 1994).

BACs largely aleviate the concerns associate with plasmids. The advantages come with disadvantages; they are slow to clone and PCR based experiments must be set up carefully to distinguish endogenous and BAC sequence. For this reason we concentrated on larger deletions once the functional binding site mutations yielded ambiguous results. We found that deletions of nearly the entire enhancer, leaving one or two fragment E CpGs, was sufficient to remove low methylation activity from the area. However, no other deletion could obviate the presense of low enhancer methylation, even in knockout serum. The largest effect we saw was 20-40% increases in methylation, which suggests that the window is maintained by cooperative binding along the 1kb enhancer region. Large deletions may remove some trancription factor binding, but the

remaining CpGs still have local binding events which protect them from DNA methylation. This complicates the study of the *Il12b* enhancer window, as DNA methylation cannot be targeted for deletion in a large context. Counterintuitively, the best approach to dissecting the multiple binding sites at the *Il12b* enhancer may be to analyze smaller regions inserted into a different sequence context.

Complicating analysis of the enhancer deletion BACs was the fact that *Il12b* enhancer methylation is extremely variable in ES cells, in contrast to the extremely low and stable methylation found in almost every other cell type assayed. Notably, the highest level of *Il12b* enhancer methylation in our study was at the formation of embryoid bodies, an event that simulates germ layer formation during embryogenesis. During this stage the *Il12b* enhancer window approaches genomic background levels in a large proportion of the differentiating cells. The high methylation at the enhancer window may be related to the balance between pluripotency and differentiation. Even within pluripotent lines, different backgrounds have different *Il12b* enhancer methylation, perhaps related to the efficiency of their pluripotency network (Skottman et al., 2005).

Changing ES media to utilize knockout serum had a large effect on *Il12b* enhancer methylation. One study suggests knockout serum drives lower methylation in ES cells via the increased Vitamin C content, which may effect the Tet hydroxylase pathway (Chung et al., 2010). Another possibility is that using defined knockout serum may remove contamination with negative regulators of pluripotency. Lower *Il12b* enhancer methylation could possibly be acheived in 2i+LIF media, which currently is the best described culture condition for ES cells (Wray et al., 2011).

Finally, we find that the methylation state at the *Il12b* enhancer does not strongly correlate with expression in embryonic stem cell derived macrophages. Despite the extremely low methylation of bone marrow derived macrophages, ES derived macrophages still have moderate methylation at the *Il12b* enhancer. The level of *Il12b* enhancer DNA methylation was not predictive of expression intensity in the lines tested. It is possible that the mixed methylation at the *Il12b* enhancer is representative of a diverse cell population, where some cells have very low methylation and are responsible for the majority of expression. Another possibility is that that transcription factor binding which results in only moderate protection from DNA methylation is still functional for enhancing transcription. This hypothesis is based on the consideration that chromatin modification at the *Il12b* enhancer is likely a minor effect of pioneer transcription factor binding, as the *Il12b* enhancer remains generally inaccessible until just before expression. In this model, transcription factors with weak demethylation activity likely require developmental time on the order of weeks establish a low methylation window at the *Il12b* enhancer post differentiation, while in culture the ES derived macrophages only get 7 days.

Despite the ambiguity of targeted deletions in the *Il12b* enhancer, we clearly see that this region is regulated well before lineage commitment, by changes in media conditions and by differentiation. Untangling the logic of the diverse phenomena at the *Il12b* enhancer will increase our understanding of pioneer factor binding during pluripotency for tissue specific genes.

Materials and Methods

Cell culture and reagents

The R1, CCE, and IPS 1-A2 murine ES lines were grown in Knockout DMEM supplemented with 15% fetal bovine serum (Omega), 0.1 mM nonessential amino acids, 2 mM L-glutamine, 1% penicillin/streptomycin, 0.05 mM β-mecaptoenthanol, and 1000 U/ml LIF (ESGRO, Millipore). Defined media was the same except fetal bovine serum was replaced with 15% KnockOut^{TM} SR. All culture products were purchased from Gibco unless otherwise noted. ES cells were maintained in gelatin (Stem Cell Technologies) coated Petri dishes and on a layer of mouse embryonic fibroblasts mitotically inactivated with mytomycin-C, where appropriate. ES cells were removed from plates using Trypsin-EDTA (Stem Cell Technologies), treated for 5 minutes and then neutralized by FBS. The HoxB8 line was maintained as previously described (Wang et al., 2006).

BAC Modification and Preparation

The *Il12b* BAC was purchased from CHORI-BACPAC and modified by insertion of a GFP cassette into the second exon (Pope, unpublished data). Insertion of exogenous sequence into the BAC was done according to a protocol adapted from (Gong and Yang, 2005). BACs were electroporated into SW102 RecA expressing bacteria and selected for targeted recombination of GalK and replacement of GalK by minimal galactose media or deoxygalactose respectively (Warming et al., 2005). For stable ES cell transduction, a PGK-Neomycin expressing cassette was introduced into the BAC as described in (Wang, 2001). Successful recombineering was confirmed by restriction enzyme fingerprinting and sequencing of the insert region.

To prepare for ES cell transduction, BAC DNA was isolated using the Large Construct Kit (Qiagen) and linearized with the restriction enzyme PI-SceI. Pre-methylation of BACs was done by overnight incubation with *Sss*I methylase and SAM. BAC DNA was then phenol chloroform extracted and resuspended in 500uL PBS for electroporation. BAC integrity was verified on a large pulse field gel (BIO-Rad CHEF Mapper XA).

Generation of Stable ES Cell BAC lines

ES cells were grown to confluency in a 10cm plate prior to transduction with 10ug of plasmid DNA using Lipafectamine (Invitrogen), or with 5-20ug of BAC DNA by electroporation at 0.27kV 500uFd. After a short recovery, ES cells were replated 1:2. Selection for plasmid or BAC integration was done using the antibiotic G418/Neomycin at 255ug/ul for approximately ten days. At this point single colonies were picked and outgrown into stable clones, maintained in G418. Genomic DNA was isolated from stable ES clones with the DNeasy kit (Qiagen). Integration of plasmid or BAC DNA was confirmed by genotyping PCR.

Bisulfite Sequencing

Bisulfite treatment of 1-2.5ug of genomic DNA was performed overnight at 55 C, following denaturation by 5ul of 3M NaOH. The bisulfite-treated DNA was desalted using the PCR Purification kit (Qiagen), then was neutralized with 5m ammonium acetate and precipitated with 2mg yeast tRNA. Bisulfite-treated DNA was resuspended in 50ul TE.

Sequence-specific PCR of the bisulfite-treated DNA was performed using primers specific to primer or BAC regions. The PCR fragments were cloned into the pCRII vector (Invitrogen, K2070-20) and transformed into DH5α *E. coli* cells. Miniprep plasmid DNA was sequenced using M13 reverse primers.

ES Derived Macrophage Differentiation

The embryonic stem cell derived macrophage protocol was adapted from previous work (Keller et al., 1993, 2002)(Scott Pope, unpublished data). Briefly, ES cells are induced to differentiate by removal of LIF and growth in IMDM media. Cells are resuspended in Embryoid Body media which consists of IMDM base media (Cell-gro), 15% general lab FBS (Omega Scientific), 0.4mM Monothioglycerol (Sigma), 1% pen/strep, 2% L-glutamine, 300 μg/ml transferrin (Roche), 50 μg/ml Ascorbic Acid (Sigma), and 5% Protein-Free Hybridoma Medium (PFHM-II, Gibco). After six days, embryoid bodies are physcially disrupted and grown in Macrophage media 1 consisting of IMDM base media , 10% general lab FBS, 0.15mM Monothioglycerol, 1% pen/strep, 1% L-glutamine, 5% CMG media (M-CSF conditioned media) and 1 ng/ml IL-3. After 48 hours, suspension cells were transferred to Macrophage media without IL-3, and grown for five days before stimulation with 100ng/ml Lipopolysacchride.

FACS

Staining of EB cells was performed with a standard protocol with antibodies for CKit and CD41. Cells were assayed and sorted on a FACS AriaII. ES derived macrophages were also confirmed with flow cytometry by staining with f4/80 and CD11b on a FACSCalibur.

RT-qPCR

RNA was extracted using Tri-Reagent (MRC) and purified with the RNeasy kit (Qiagen). cDNA was prepared from 1ug of RNA using the Omniscript RT Kit (Qiagen) primed with

random hexamers. cDNA was diluted 1:5 and analyzed by qPCR on an iCycler (BioRad). *Il12b, Il-6* and *RANTES* primers were previously described (Ramirez-Carrozzi et al., 2009).

Methylome Computational Display

All genome wide bisulfite sequencing methylomes were displayed on UCSC Genome Browser in the mm9 build. The ES cell methylome was done in ROSA V6.5 ES cells (Kathrin Plath lab, unpublished data), and the macrophage data was done in BL6 peritoneal macrophages (Lusis lab, unpublished data). The mouse Frontal Cortex methylome was obtained from publically available data (Lister et al., 2013).

Figure Legends

Figure 2-1 – Large Portions of the *Il12b* Enhancer Region Can Trigger Low Methylation (A) Schematic of the plasmid vectors with *Il12b* enhancer fragments inserted before the *Il12b* promoter and the Red Flourescent Protein gene. CpGs in the plasmid sequence are shown as ball and stick figures. (B) Bisulfite sequencing of the *Il12b* C Enhancer plasmid stably integrated in ES cells. CpGs are shown on the left, numbered based on sequence order. Across is the methylation status for that CpG in each condition shown at top. Methylation status is displayed as % of methylated CpGs versus total CpGs and the Ratio is the methylated CpGs to total CpGs sequenced. Each CpG is colored according to methylation status, see legend. Premethylation for the last three clones occurred *in vitro* with SssI prior to transfection. (C) Bisulfite sequencing of the *Il12b* CE Enhancer plasmid stably integrated in ES cells. For the Enhancer CpGs, position on left is their endogenous distance from the transcription start site of *Il12b*. Note that all clones received pre-methylated plasmid (D) Bisulfite sequencing of the *Il12b* DE Enhancer plasmid stably integrated in ES cells.

Figure 2-2 – Integrated Plasmids are Susceptible to Changes in Methylation State (A) Bisulfite sequencing of the *Il12b* promoter only plasmid stably integrated in ES cells (B) Bisulfite sequencing of the hCMV enhancer *Il12b* promoter plasmid. The first column contains the bisulfite sequencing of the pre-methylated plasmid alone before transfection.

Figure 2-3 – BAC Deletion Mutations Targeting the Low Methylation at the *Il12b* Enhancer (A) Schematic of the BAC modifications to the 191kb *Il12b* BAC, at the *Il12b*

enhancer (B) Bisulfite sequencing of the *Il12b* Enhancer Del BAC stably transfected into ES cells. Grey represents CpGs deleted in the BAC. The *Il12b* Enhancer Up and Down regions are immediately adjacent to the *Il12b* CE enhancer fragment and represent the closest CpGs. (C) Bisulfite sequencing of the C/EBP Mutant Enhancer *Il12b* BAC stably transfected into ES cells. Bisulfite sequencing results from the endogenous *Il12b* enhancer locus in each clone is pooled in the first column. (D) Bisulfite sequencing of the Oct Mutant Enhancer *Il12b* BAC stably transfected into ES cells. The New CG position represents a CpG introduced by substitution mutation at the Oct binding site.

Figure 2-4–Binding Site Mutations and Large Deletions Within the *Il12b* Enhancer Cannot Remove the Low Methylation Window

(A) Bisulfite sequencing of the Evi-1 Enhancer Deletion *Il12b* BAC stably transfected into ES cells. The first column contains bisulfite sequencing of the pre-methylated BAC alone, followed by the endogenous enhancer locus for all four clones, followed by the BAC locus for all four clones. (B) Bisulfite sequencing of the Half-DE Enhancer Deletion *Il12b* BAC stably transfected into ES cells. Deleted CpGs are colored grey. (C) Bisulfite sequencing of the Half-DE 2 Enhancer Deletion *Il12b* BAC stably transfected into ES cells.

Figure 2-5– Embryonic Stem Cells Have Uniquely Variable Methylation at the *Il12b* Enhancer in Contrast to Primary Cells

(A) Bisulfite sequencing of several wildtype cell lines at the endogenous *Il12b* enhancer. Bone Marrow Derived Macrophages were cultured from BL6, HoxB8s are a transgene driven macrophage progenitor line, J774 is a transformed macrophage cell line, and embryoid bodies

are ES cells triggered to differentiate by withdrawl of LIF from culture media. (B) Bisulfite sequencing at the *Il12b* enhancer of ES lines with variable MEF feeder inclusion. (C) Bisulfite sequencing at the *Il12b* enhancer of R1 ES lines in various growth conditions. No Feeders indicated ES cells were plated without a MEF feeder layer and grown for 1 day. Overgrown indicates ES cell DNA was isolated only at extremely high confluency for bisulfite sequencing. (D) Bisulfite sequencing of the *Il12b* enhancer in feeder independent CCE line with increasing amounts of β-mercaptoethanol supplemented into the culturing media. 2x and 4x represents fold over usual amount, given in Methods. (E) The *Il12b* enhancer visualized in UCSC Genome Browser, with methylome data from Peritoneal Macrophages, Frontal Cortex, and ES cells. Each vertical line represents a CpG. The height of each line represents the percent of methylation at that locus. The location of the CE fragment of the *Il12b* enhancer is shown below.

Figure 2-6 – Evaluating *Il12b* Enhancer Mutations in Knockout Serum Media

(A) Bisulfite sequencing of the *Il12b* enhancer in ES cells grown in Fetal Bovine Serum or in Knockout Serum (B) Bisulfite sequencing of stable Half-DE enhancer deletion *Il12b* BAC ES lines grown in knockout serum. (C) Bisulfite sequencing of stable Half-DE 2 enhancer deletion *Il12b* BAC ES lines grown in knockout serum.

Figure 2-7 – Large Scale Deletions of the *Il12b* Enhancer Window Demonstrate the Irreducability of the Low Methylation Window

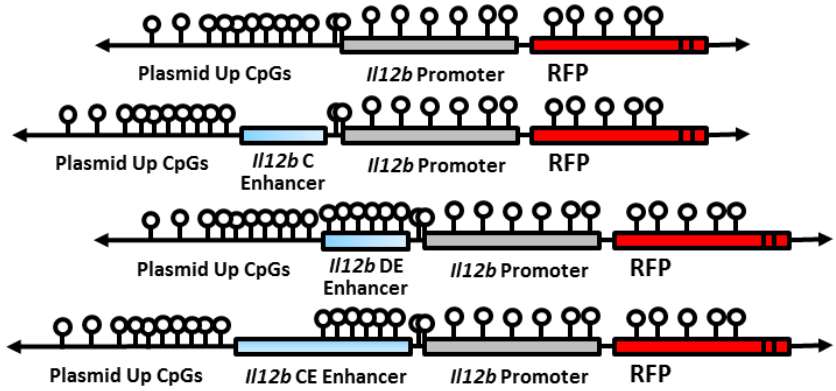
(A) Schematic of the modifications to the *Il12b* BAC involving deletion of the C enhancer fragment (B) Bisulfite sequencing of the CD Enhancer Deletion *Il12b* BAC stably transfected

into ES cells. (C) Bisulfite sequencing of the C-Half DE Enhancer Deletion *Il12b* BAC stably transfected into ES cells.

Figure 2-8 – ES lines Differentiated into Macrophages Experience Small Changes at the *Il12b* Enhancer Window (A) Bisulfite sequencing of R1 ES cells, ROSA V6.5 stem cells, and induced pluripotent stem cell line 1-A2 at the pluripotent stage. (B) Bisulfite sequencing of R1 ES cells, ROSA V6.5 stem cells, and IPS cells after differentiation into macrophages. (C) RT-qPCR for three inflammatory genes, including *Il12b,* expressed in response to LPS stimulation for 2 hours. Magnitude of expression is shown as fold over unstimulated, GAPDH adjusted. Error bars represent the standard error from two replicates per cell line. (D) Bisulfite sequencing of the *Il12b* enhancer in R1 cells isolated at different times during an in vitro differentiation into macrophages. At the EB stage cells were separated by FAC sorting for Ckit staining or Ckit CD41 double positive staining. At the MacI stage cells could be easily separated by taking the supernatant. ES macrophage identity was confirmed by flow cytometry for Cd11b and F480 (data not shown).

Figure 2-1 – Large Portions of the *Il12b* Enhancer Region Can Trigger Low Methylation

A



B

		R1 ES cells with Il12b C Enhancer + Promoter Plasmid											
		Clone 1		Clone 2		Clone 3		Pre-Methylated Clone 1		Pre-Methylated Clone 2		Pre-Methylated Clone 3	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Plasmid Up	1	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	27	3/11
	2	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	36	4/11
	3	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	36	4/11
	4	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	45	5/11
	5	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	45	5/11
	6	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	45	5/11
	7	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	36	4/11
	8	0	0/9	0	0/9	13	1/8	0	0/10	0	0/11	27	3/11
	9	0	0/9	0	0/9	0	0/8	0	0/10	0	0/11	0	0/11
Plasmid Il12b Promoter	1	0	0/10	0	0/12	9	1/11	0	0/12	18	2/11	0	0/11
	2	0	0/11	0	0/12	9	1/11	0	0/12	18	2/11	0	0/11
	3	0	0/10	0	0/12	9	1/11	0	0/12	27	3/11	0	0/11
	4	0	0/10	0	0/12	9	1/11	0	0/12	27	3/11	0	0/10
	5	0	0/9	0	0/12	9	1/11	8	1/12	27	3/11	0	0/10

% of CpGs Methylated
0
20
40
60
80+

C

		R1 ES Cells with Pre-methylated *Il12b* CE Enhancer and Promoter Plasmid					
		Clone 1		Clone 2		Clone 3	
	CpG	%	Ratio	%	Ratio	%	Ratio
Plasmid Il12b Enhancer	-9874	14	1/7	56	5/9	0	0/7
	-9777	20	1/5	0	1/8	17	2/8
	-9646	17	1/6	50	5/10	0	0/10
	-9617	0	0/6	10	1/10	13	1/8
	-9512	0	0/6	0	0/11	13	1/8
	-9420	17	1/6	40	4/10	10	1/10
Plasmid Il12b Promoter	1	50	4/8	100	11/11	50	5/10
	2	25	2/8	91	10/11	11	1/9
	3	25	2/8	64	7/11	30	3/10
	4	38	3/8	33	4/12	33	3/9
	5	25	2/8	30	3/10	33	3/9
	6	29	2/7	83	10/12	44	4/9
	7	29	2/7	42	5/12	0	0/9

D

		R1 ES Cells with Pre-methylated *Il12b* DE Enhancer and Promoter Plasmid									
		Clone 1		Clone 2		Clone 3		Clone 4		Clone 5	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Plasmid Il12b Enhancer	-9874	0	0/9	33	2/6	100	5/5				
	-9777	0	0/10	44	4/9	86	6/7	18	2/11	57	4/7
	-9646	0	0/13	25	2/8	86	6/7	17	2/12	63	5/8
	-9617	0	0/12	22	2/9	86	6/7	17	2/12	63	5/8
	-9512	21	3/14	20	2/10	86	6/7	15	2/13	50	4/8
	-9420	29	4/14	20	2/10	86	6/7	29	4/14	13	1/8
Plasmid Il12b Promoter	1	57	8/14	20	2/10	100	7/7	57	8/14	63	5/8
	2	14	2/14	20	2/10	100	7/7	21	3/14	38	3/8
	3	57	8/14	20	2/10	86	6/7	21	3/14	63	5/8
	4	43	6/14	30	3/10	86	6/7	21	3/14	63	5/8
	5	14	2/14	0	0/10	86	6/7	43	6/14	25	2/8
	6	7	1/14	20	2/10	86	6/7	36	5/14	63	5/8
	7	7	1/14	0	0/10	83	5/6	14	2/14	38	3/8

Figure 2-2 - Integrated Plasmids are Susceptible to Changes in Methylation State

A

		R1 ES cells with No Enhancer + Il12b Promoter Plasmid															
		Clone 1		Clone 2		Clone 3		Pre-Methylated Clone 1		Pre-Methylated Clone 2		Pre-Methylated Clone 3		Pre-Methylated Clone 4		Pre-Methylated Clone 5	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Plasmid Up	1	80	8/10	73	8/11	90	9/10	100	14/14	92	12/13	90	11/13	0	0/13	40	4/10
	2	82	9/11	73	8/11	70	7/10	77	10/13	100	12/12	92	12/13	0	0/14	55	6/11
	3	82	9/11	89	8/9	100	10/10	85	11/13	79	11/14	100	13/14	7	1/14	64	7/11
	4	90	9/10	83	5/6	100	10/10	100	13/13	86	12/14	100	13/14	23	3/13	50	5/10
	5	90	9/10	83	5/6	100	10/10	85	11/13	92	12/13	100	14/14	29	4/14	55	6/11
	6	90	9/10	43	3/7	100	10/10	92	12/13	85	11/13	100	12/13	21	3/14	45	5/11
	7	90	9/10	100	6/6	100	10/10	100	12/12	100	12/12	100	14/14	7	1/14	45	5/11
	8	90	9/10	100	6/6	100	10/10	92	12/13	100	12/12	100	14/14	7	1/14	45	5/11
Plasmid Il12b Promoter	9	90	9/10	100	6/6	100	10/10	100	14/14	100	13/13	100	13/13	7	1/14	36	4/11
	10	78	7/9	100	7/7	100	10/10	100	14/14	85	11/13	100	13/13	21	3/14	55	6/11
	11	45	5/11	75	6/8	90	9/10	79	11/14	100	13/13	90	14/14	0	0/14	45	5/11
	12	90	9/10	86	6/7	80	8/10	69	9/13	79	11/14	80	13/14	29	4/14	55	6/11
	13	90	9/10	86	6/7	100	10/10	100	13/13	92	12/13	100	14/14	21	3/14	55	6/11

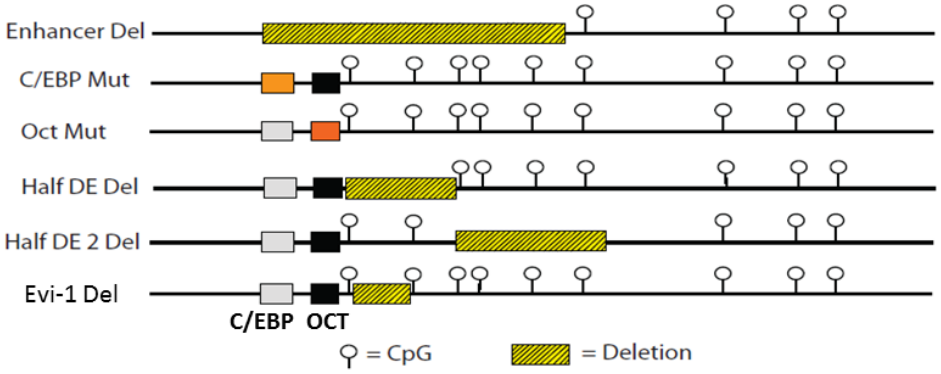
% of CpGs Methylated
0
20
40
60
80+

B

		R1 ES cells with Pre-Methylated hCMV-Enhancer + Il12b Promoter Plasmid					
		Pre Methylated Plasmid		Clone 1		Clone 2	
	CpG	%	Ratio	%	Ratio	%	Ratio
hCMV Enhancer	1	100	10/10	0	0/9	0	0/7
	2	100	10/10	0	0/9	0	0/7
	3	100	10/10	0	0/9	0	0/7
	4	100	10/10	0	0/9	0	0/7
	5	100	10/10	0	0/9	0	0/7
	6	100	10/10	0	0/9	0	0/7
	7	100	10/10	0	0/9	14	1/7
	8	100	10/10	0	0/9	0	0/7
	9	100	10/10	22	2/9	0	0/7
	10	100	10/10	0	0/9	0	0/7
	11	100	10/10	0	0/9	0	0/7
Plasmid Il12b Promoter	1			0	0/9	0	0/7
	2			0	0/9	0	0/7
	3			0	0/9	0	0/7
	4			0	0/9	0	0/7
	5			11	1/9	0	0/7
	6			0	0/9	0	0/7
	7			0	0/9	0	0/7

Figure 2-3 – BAC Deletion Mutations Targeting the Low Methylation at the *Il12b* Enhancer

A



B

		R1 ES cells with *Il12B* Enhancer Deletion BAC					
		Clone 1		Clone 2		Clone 3	
	CpG	%	Ratio	%	Ratio	%	Ratio
Il12b Enh Up	-10328	57	8/14	87	13/15	33	5/15
	-10272	93	14/15	87	13/15	93	14/15
Il12b Enhancer	-9874						
	-9777						
	-9646						
	-9617						
	-9512						
	-9420	62	18/29	73	22/30	47	7/15
Il12b Enh Down	-9161	78	7/9	79	11/14		

% of CpGs Methylated
0
20
40
60
80+

C

		ES Cells with CEBP Mutant Enhancer *Il12b* BAC							
		Endogenous Enhancer Pooled		Clone 1 BAC		Clone 2 BAC		Clone 3 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	78	14/18	17	3/18	86	18/21	15	2/13
	-9777	53	8/15	94	17/18	95	19/20	82	9/11
	-9646	75	9/12	43	3/7	50	4/8	50	1/2
	-9617	40	4/10	57	4/7	38	3/8	50	1/2

D

		ES Cells with Oct Mutant Enhancer *Il12b* BAC					
		Endogenous Enhancer Pooled		Clone 1 BAC		Clone 2 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	New CG			100	6/6	100	13/13
	-9874	30	7/23	67	4/6	8	1/13
	-9777	68	13/19	100	4/4	50	6/12
	-9646	47	8/17	67	2/3	8	1/12
	-9617	41	7/17	0	0/3	42	5/12

Figure 2-4 – Binding Site Mutations and Large Deletions Within the *Il12b* Enhancer Cannot Remove the Low Methylation Window

A

		ES cells with Pre-Methylated Evi-1 Enhancer Deletion *Il12b* BAC															
		Pre-Methylated BAC		Clone 1 Endogenous		Clone 2 Endogenous		Clone 3 Endogenous		Clone 4 Endogenous		Clone 1 BAC		Clone 2 BAC		Clone 3 BAC		Clone 4 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	75	6/8	42	5/12	54	7/13	68	13/19	40	2/5	13	2/16	14	2/14	14	2/14	28	5/18
	-9777	71	5/7	40	4/10	85	11/13	76	13/17	80	4/5	82	14/17	50	7/14	50	7/14	72	13/18
	-9646	63	5/8	8	1/12	54	7/13	53	10/18	20	1/5	43	6/14	50	4/8	50	4/8	47	8/17
	-9617	88	7/8	42	5/12	33	3/9	35	7/20	40	2/5	43	6/14	38	3/8	38	3/8	38	6/16

B

		ES cells with Pre-Methylated Half DE Enhancer Deletion *Il12b* BAC									
		Pre-Methylated BAC		Clone 1 Endogenous		Clone 2 Endogenous		Clone 1 BAC		Clone 2 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874			38	8/21	7	1/14				
	-9777			67	14/21	36	5/14				
	-9646	100	11/11	57	12/21	14	2/14	27	4/15	15	2/13
	-9617	100	11/11	43	9/21	0	0/14	80	12/15	38	5/13
	-9512	100	11/11					93	14/15	43	6/14
	-9420	90	9/10					67	10/15	20	3/15

% of CpGs Methylated
0
20
40
60
80+

C

		ES cells with Pre-Methylated Half DE 2 Enhancer Deletion *Il12b* BAC															
		Pre-Methylated BAC		Clone 1 Endogenous		Clone 2 Endogenous		Clone 3 Endogenous		Clone 4 Endogenous		Clone 1 BAC		Clone 2 BAC		Clone 3 BAC		Clone 4 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	100	20/20	26	5/19	25	3/12	19	3/16	50	7/14	35	7/20	20	4/20	15	3/20	53	9/17
	-9777	100	20/20	42	8/19	75	9/12	56	9/16	50	7/14	55	11/20	63	12/19	80	16/20	82	14/17
	-9646			22	4/18	18	2/11	38	6/16	79	11/14								
	-9617			33	6/18	50	6/12	38	6/16	57	8/14								
Enhancer Down	-9161											70	7/10	78	7/9	70	7/10	60	6/10
	-8953											82	9/11	67	6/9	50	5/10	80	8/10
	-8882											80	8/10	100	9/9	50	5/10	80	8/10

Figure 2-5 – Embryonic Stem Cells Have Uniquely Variable Methylation at the *Il12b* Enhancer in Contrast to Primary Cells

A

		Bone Marrow Derived Macrophages		HoxB8 Macrophage Progenitors		J774 Macrophages		R1 ES cells		Embryoid Bodies (RI)	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enh Up	-10501	100	6/6	83	5/6	100	6/6	67	4/6	71	5/7
	-10328	50	3/6	17	1/6	83	5/6	83	5/6	90	9/10
	-10272	100	6/6	83	5/6	100	6/6	83	5/6	100	10/10
Il12b Enhancer	-9874	8	1/12	0	0/12	8	1/12	42	5/12	53	8/15
	-9777	25	3/12	8	1/12	8	1/12	58	7/12	87	13/15
	-9646	0	0/12	0	0/10	8	1/12	50	6/12	75	9/12
	-9617	0	0/12	0	0/10	0	0/11	42	5/12	60	9/15

% of CpGs Methylated
0
20
40
60
80+

B

		Bone Marrow Derived Macrophages		CCE ES cells		R1 ES cells, Confluent		R1 ES Cells Feeder Depleted		MEF wt Feeder Cells	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enh Up	-10501	75	6/8	20	2/10	60	6/10	80	8/10	89	8/9
	-10328	67	6/9	40	4/10	56	5/9	80	8/10	50	5/10
	-10272	89	8/9	90	9/10	60	6/10	80	8/10	90	9/10
Il12b Enhancer	-9874	0	0/11	18	2/11	18	2/11	17	2/12	10	1/10
	-9777	18	2/11	73	8/11	55	6/11	77	10/13	9	1/11
	-9646	0	0/10	20	2/10	20	2/10	38	5/13	0	0/11
	-9617	10	1/10	50	5/10	20	2/10	23	3/13	0	0/11
Il12b Enhancer Down	-9161	78	7/9	70	7/10	40	4/10	86	6/7	56	5/9
	-8953	100	9/9	90	9/10	50	5/10	78	7/9	89	8/9
	-8882	100	9/9	50	5/10	90	9/10	78	7/9	78	7/9

C

		R1 wt with Feeders		R1 wt No Feeders		R1 wt Overgrown	
	CpG	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	45	5/11	8	1/12	25	3/12
	-9777	64	7/11	50	6/12	58	7/12
	-9646	40	2/5	55	6/11	33	3/9
	-9617	60	3/5	58	7/12	20	2/10

D

		CCE ES Cells					
		1x BME		2x BME		4x BME	
	CpG	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	11	1/9	11	1/9	13	1/8
	-9777	56	5/9	78	7/9	57	4/7
	-9646	11	1/9	44	4/9	25	2/8
	-9617	56	5/9	33	3/9	50	4/8

E

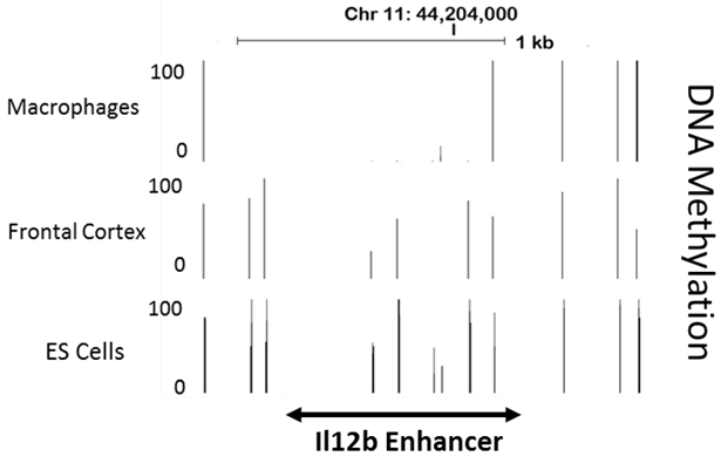


Figure 2-6 – Evaluating *Il12b* Enhancer Mutations in Knockout Serum Media

A

		R1 ES cells			
		ES FBS		ES Knock-Out Serum	
	CpG	%	Ratio	%	Ratio
Il12b Enhancer	-9874	38	5/13	13	2/15
	-9777	55	6/11	67	8/12
	-9646	45	5/11	25	3/12
	-9617	55	6/11	25	3/12

% of CpGs Methylated
0
20
40
60
80+

B

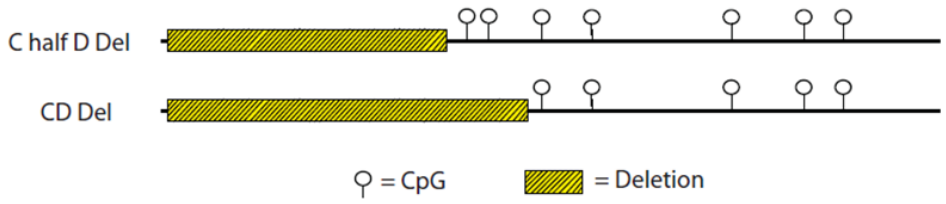
		ES cells with Pre-Methylated Half DE Enhancer Deletion *Il12b* BAC, in KSR Media									
		Pre-Methylated BAC		Clone 1 Endogenous		Clone 2 Endogenous		Clone 1 BAC		Clone 2 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874			13	1/8	38	3/8				
	-9777			38	3/8	63	5/8				
	-9646	100	11/11	0	0/5	25	2/8	75	3/4	38	3/8
	-9617	100	11/11	40	2/5	38	3/8	50	2/4	63	5/8
	-9512	100	11/11					100	4/4	100	8/8
	-9420	90	9/10					50	2/4	75	6/8

C

		ES cells with Pre-Methylated Half DE 2 Enhancer Deletion *Il12b* BAC, in KSR Media									
		Pre-Methylated BAC		Clone 1 Endogenous		Clone 2 Endogenous		Clone 1 BAC		Clone 2 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	100	20/20	13	1/8	13	1/8	18	3/17	13	1/8
	-9777	100	20/20	25	2/8	13	1/8	59	10/17	50	4/8
	-9646			38	3/8	0	0/5				
	-9617			50	4/8	0	0/5				
	-9512										
	-9420										

Figure 2-7 – Large Scale Deletions of the *Il12b* Enhancer Window Demonstrate the Irreducibility of the Low Methylation Window

A



B

		ES cells with Pre-Methylated CD Enhancer Deletion Il12b BAC							
		Clone 1 Endogenous		Clone 2 Endogenous		Clone 1 BAC		Clone 2 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enh Up	-10501								
	-10328							79	11/14
	-10272					100	11/11	93	13/14
Il12b Enhancer	-9874	55	6/11	8	1/12				
	-9777	40	4/10	50	6/12				
	-9646			27	3/11				
	-9617			45	5/11				
Enh Down	-9512					82	9/11	79	11/14
	-9420					75	9/12	71	10/14

% of CpGs Methylated
0
20
40
60
80+

C

		ES cells with Pre-Methylated C-Half DE Enhancer Deletion *Il12b* BAC															
		Pre-Methylated BAC		Clone 1 Endogenous		Clone 2 Endogenous		Clone 3 Endogenous		Clone 4 Endogenous		Clone 1 BAC		Clone 2 BAC		Clone 3 BAC		Clone 4 BAC	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enh Up	-9161																		
	-8953																		
	-8882											83	10/12	58	7/12	78	7/9	82	9/11
Il12b Enhancer	-9874			44	4/9	14	1/7	11	1/9	38	3/8								
	-9777			44	4/9	14	1/7	83	5/6	57	4/7								
	-9646	100	8/8	22	2/9	43	3/7	17	1/6	43	3/7	25	3/12	33	4/12	33	3/9	36	4/11
	-9617	100	8/8	33	3/9	14	1/7	0	0/6	14	1/7	58	7/12	58	7/12	30	3/10	18	2/11
	-9512											33	4/12	33	4/12	50	5/10	70	7/10
	-9420											83	10/12	50	6/12	60	6/10	82	9/11

Figure 2-8 – ES lines Differentiated into Macrophages Experience Small Changes at the *Il12b* Enhancer Window

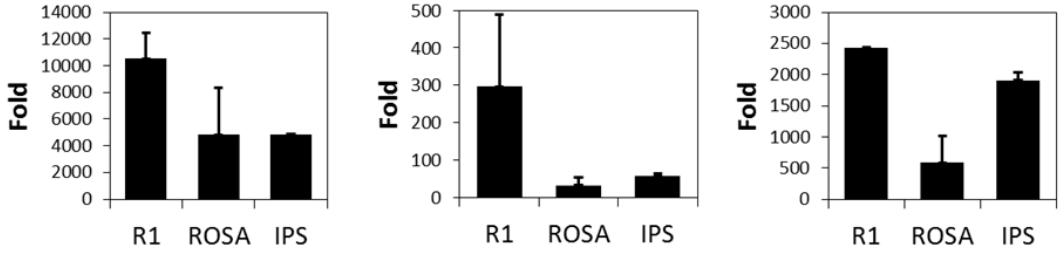
A

		Embryonic Stem Cell Lines - ES Stage					
		R1		ROSA		IPS	
	CpG	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	43	13/30	83	19/23	21	5/24
	-9777	70	21/30	91	21/23	38	9/24
	-9646	53	16/30	78	18/23	21	5/24
	-9617	43	13/30	74	17/23	29	7/24

B

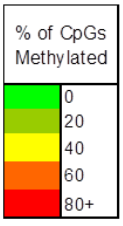
		Embryonic Stem Cell Lines - Differentiated to Macrophage Stage					
		R1		ROSA		IPS	
	CpG	%	Ratio	%	Ratio	%	Ratio
Il12b Enhancer	-9874	33	8/24	19	4/21	5	1/21
	-9777	54	13/24	59	13/22	9	2/22
	-9646	67	16/24	75	18/24	9	2/23
	-9617	58	14/24	43	10/23	0	0/23

C



D

		R1 ES Cells		R1 LIF Withdrawal		R1 EB Day 3		R1 EB Day 3 Ckit sorted		R1 EB Day 6		R1 EB Day 6 Ckit-CD41 Sorted		R1 MacI Suspension		R1 Mac1 Attached		R1 Macro-phages	
	CpG	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b Enh Up	-10501	71	5/7	88	7/8	88	7/8	63	5/8	88	7/8	75	6/8	44	4/9	60	6/10	90	9/10
	-10328	86	6/7	100	8/8	88	7/8	75	6/8	100	8/8	100	8/8	67	6/9	70	7/10	100	10/10
	-10272	100	7/7	88	7/8	88	7/8	100	8/8	100	8/8	100	8/8	67	6/9	100	10/10	100	10/10
Il12b Enhancer	-9874	62	8/13	42	5/12	33	4/12	56	10/18	71	10/14	38	3/8	38	3/8	67	6/9	44	4/9
	-9777	77	10/13	75	9/12	92	11/12	67	12/18	79	11/14	50	5/10	78	7/9	78	7/9	70	7/10
	-9646	69	9/13	43	6/14	85	11/13	78	14/18	56	9/16	58	7/12	75	6/8	75	6/8	43	6/14
	-9617	38	5/13	64	9/14	77	10/13	78	14/18	44	7/16	17	2/12	56	5/9	78	7/9	43	6/14



References

Allshire, R.C., Javerzat, J.P., Redhead, N.J., and Cranston, G. (1994). Position effect variegation at fission yeast centromeres. Cell *76*, 157–169.

Blum, R., Vethantham, V., Bowman, C., Rudnicki, M., and Dynlacht, B.D. (2012). Genome-wide identification of enhancers in skeletal muscle: the role of MyoD1. Genes Dev. *26*, 2763–2779.

Boyer, L. a, Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. *41*, e155.

Chung, T.-L., Brena, R.M., Kolle, G., Grimmond, S.M., Berman, B.P., Laird, P.W., Pera, M.F., and Wolvetang, E.J. (2010). Vitamin C promotes widespread yet specific DNA demethylation of the epigenome in human embryonic stem cells. Stem Cells *28*, 1848–1855.

Felsenfeld, G., Boyes, J., Clark, D., and Studitsky, V. (1996). Chromatin structure and gene expression. Proc. Natl. Acad. Sci. USA *93*, 9384–9388.

Gong, S., and Yang, X.W. (2005). Modification of bacterial artificial chromosomes (BACs) and preparation of intact BAC DNA for generation of transgenic mice. Curr. Protoc. Neurosci. *Chapter 5*, Unit 5.21.

Grosveld, F., Assendelft, G. van, Greaves, D., and Kollias, G. (1987). Position-independent, high-level expression of the human β-globin gene in transgenic mice. Cell *51*, 975–985.

Heintz, N. (2001). BAC to the future: the use of bac transgenic mice for neuroscience research. Nat. Rev. Neurosci. *2*, 1–10.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K. a, et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Keller, G., Kennedy, M., Papayannopoulou, T., and Wiles, M. V (1993). Hematopoietic commitment during embryonic stem cell differentiation in culture. Mol. Cell. Biol. *13*, 473–486.

Keller, G.M., Webb, S., and Kennedy, M. (2002). Hematopoietic Development of ES Cells in Culture. Methods Mol. Med. *63*, 209–230.

Li, Q., Peterson, K.R., Fang, X., and Stamatoyannopoulos, G. (2002). Locus control regions. Blood *100*, 3077–3086.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Mukamel, E. a, Nery, J.R., Urich, M., Puddifoot, C. a, Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science (80-.). *341*, 1237905.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet. *38*, 431–440.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Moore, K.J., Fabunmi, R.P., Andersson, L.P., and Freeman, M.W. (1998). In Vitro Differentiated Embryonic Stem Cell Macrophages : A Model System for Studying Atherosclerosis-Associated Macrophage Functions. Arterioscler. Thromb. Vasc. Biol. *18*, 1647–1654.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat. Biotechnol. *26*, 101–106.

Ramirez-Carrozzi, V.R., Nazarian, A. a, Li, C.C., Gore, S.L., Sridharan, R., Imbalzano, A.N., and Smale, S.T. (2006). Selective and antagonistic functions of SWI/SNF and Mi-2beta nucleosome remodeling complexes during an inflammatory response. Genes Dev. *20*, 282–296.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Samstein, R.M., Arvey, A., Josefowicz, S.Z., Peng, X., Reynolds, A., Sandstrom, R., Neph, S., Sabo, P., Kim, J.M., Liao, W., et al. (2012). Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell *151*, 153–166.

Skottman, H., Mikkola, M., Lundin, K., Olsson, C., Strömberg, A.-M., Tuuri, T., Otonkoski, T., Hovatta, O., and Lahesmaa, R. (2005). Gene expression signatures of seven individual human embryonic stem cell lines. Stem Cells *23*, 1343–1356.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Wang, Z. (2001). An Efficient Method for High-Fidelity BAC/PAC Retrofitting with a Selectable Marker for Mammalian Cell Transfection. Genome Res. *11*, 137–142.

Wang, G., Calvo, K., and Pasillas, M. (2006). Quantitative production of macrophages or neutrophils ex vivo using conditional Hoxb8. Nat. Methods *3*.

Warming, S., Costantino, N., Court, D.L., Jenkins, N. a, and Copeland, N.G. (2005). Simple and highly efficient BAC recombineering using galK selection. Nucleic Acids Res. *33*, e36.

Wray, J., Kalkan, T., Gomez-Lopez, S., Eckardt, D., Cook, A., Kemler, R., and Smith, A. (2011). Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. Nat. Cell Biol. *13*, 838–845.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J. a, Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *104*, 12377–12382.

Xu, J., Watts, J. a, Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev. *23*, 2824–2838.

Zhou, L., Nazarian, A. a, Xu, J., Tantin, D., Corcoran, L.M., and Smale, S.T. (2007). An inducible enhancer required for Il12b promoter activity in an insulated chromatin environment. Mol. Cell. Biol. *27*, 2698–2712.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., et al. (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell *152*, 642–654.

Chapter 3

Analysis of the Regulatory Logic that Controls Separate CpG Island Features

Abstract

CpG islands are distinctive regulatory regions in mammalian genomes that contain euchromatin features. Exactly how these features are acquired and how they interact with each other is not clear. Here we describe a focused study looking at individual CpG rich regions in a gene desert context in an attempt to understand the determinants of the core CpG island features: low CpG methylation, high H3K4me3 deposition, and low nucleosome occupancy. We find that transcription factor binding can protect a large region from DNA methylation, but does not recruit the H3K4me3 mark in our system. Nucleosome occupancy is apparently uncoupled from other chromatin features, but is strongly affected by transcription factor binding. In contrast, H3K4me3 requires low DNA methylation, and signal increases with the size of the CpG island. Finally, we find that increasing CpG island size, but not density, seems to trigger a low DNA methylation state independent of a strong transcription factor binding site. Our evidence suggests that CpG islands have evolved a regulatory logic that is more complex that previously appreciated, but that we are beginning to understand.

Introduction

CpG islands are a crucial feature of the mammalian genome, as they are marked by unique chromatin (Bock et al., 2007) and are located at the majority of gene promoters (Davuluri et al., 2001). They are generally defined by the unusual density of cytosine-guanine dinucleotides, which are usually depleted in the genome (Gardiner-Garden and Frommer, 1987). The study of CpG islands has attempted to relate their CpG content to the chromatin features generally found there; low DNA methylation, high histone 3 lysine 4 trimethylation (H3K4me3) and low nucleosome occupancy (Bird, 1985; Fenouil et al., 2012; Mikkelsen et al., 2007) .

Several groups have reported that CpG rich DNA is sufficient to acquire CpG island properties *in vivo* (Lienert et al., 2011; Mendenhall et al., 2010; Thomson et al., 2010). CpG island DNA, such as the Nanog promoter, can autonomously acquire low DNA methylation when introduced into a targeted locus. Mutation analysis suggests that the key to this property is transcription factor binding sites which protect the locus from methylation (Lienert et al., 2011). It has been demonstrated that promoterless DNA with high CpG density can also acquire low DNA methylation in addition to high H3K4me3 (Thomson et al., 2010). Unmethylated DNA drives enrichment of H3K4me3 via recruitment of the CXXC binding protein Cfp1. Finally, independent groups introduced into mammalian cells CG rich *E. Coli* DNA, which is presumably promoterless and depleted for mammalian transcription factor binding sites (Lienert et al., 2011; Mendenhall et al., 2010). A portion of these large CpG rich sequences were able to acquire low DNA methylation and high H3K4me3. The repressive chromatin mark H3K27me3 was also found at these potential CpG island sequences. Regarding nucleosome occupancy at CpG islands, it has been demonstrated *in vitro* that CG rich DNA destabilizes nucleosome binding (Ramirez-Carrozzi et al., 2009). The current models predict that CpG rich DNA can acquire CpG island features naturally. It is unknown what the sequence requirements are to trigger this acquisition, or how the CpG island features can influence each other.

To test the model, we cloned a variety of sequences into the same system used to analyze *E. Coli* DNA in previous studies (Mendenhall et al., 2010). Novel sequences were recombineered into a bacterial artifical chromosome (BAC) containing human gene desert sequence. This provides an ideal context for studying autonomous DNA effects as the large size of BAC DNA buffers from position variegation effects (Heintz, 2001). After stable transfection of cloned BACs into mouse ES cells, we analyzed the insertion sequences by bisulfite

sequencing, chromatin immunoprecipitation, and MNase digestion. Our results indicate that transcription factors can drive local enrichment of CpG island properties, but they are not the only source of regulation. H3K4me3 is dependent on low DNA methylation, and the signal intensity of the modification is dependent on the size of the CpG rich sequence. Nucleosome occupancy intergrates signals from transcription factor binding and intrinsic DNA properties, but does not seem to influence or be influenced by DNA methylation or H3K4me3. Finally, we find evidence for a binding site independent mechanism for protection from DNA methylation, triggered by a CpG island size threshold in mouse ES cells.

Results

Dissection of the CpG Island Properties at a Small CpG Island

Initial studies discovered a small unannotated CpG island near the *Il12b* gene, a tissue specific secretory protein involved in the immune reponse. In ES cells, this region has low DNA methylation and moderate H3K4me3 signal. In order to determine whether the properties at this region are acquired autonomously, we recombineered the 270bp *Il12b* CpG Island (*Il12b* CGI) sequence into a 136 kilobase human gene desert BAC (Fig 3-2a). The BAC insertion site is unmarked by any euchromatin modifications in human cells, and does not have any significant CpG density nearby. To confirm that any establishment of low DNA methylation was caused by the cellular enviroment, we also pre-methylated all BACs *in vitro* with SssI prior to transfection. The full length insertion sequence was autonomously demethylated *in vivo* (Fig 3-1a, Fig 3-2b). To determine which portion of the *Il12b* CGI was necessary to maintain low DNA methylation, we fragmented the 270bp sequence. We found that all fragments that acquired low methylation overlapped a 6 CpG 36bp region located at 207 nucleotides into the *Il12b* CGI. The 36bp

sequence is sufficient to maintain near-zero DNA methylation. This strongly indicates the methylation status of the *Il12b* CGI is controlled by binding of a sequence specific transcription factor.

The *Il12b* CGI is also able to acquire H3K4me3 autonomously (Fig 3-1b). Chromatin immunoprecipitation shows that the *Il12b* CGI BAC insert acquires H3K4me3 nearly to the same level as the endogenous *Il12b* CGI (Fig 3-2d). Interestingly, none of the *Il12b* CGI fragments have as much H3K4me3 signal as the full length insert, despite a similar low DNA methylation status. The shorter unmethylated *Il12b* CGI fragment inserts do not acquire H3K4me3 enrichment any higher than the methylated *Il12b* fragments. The unmethylated CpGs in shorter sequences did not seem sufficient to trigger H3K4me3 deposition.

In order to determine the nucleosome occupancy at the *Il12b* CGI inserts, we utilized a qPCR based MNase protection assay. Briefly, nuclear fractions isolated from BAC transfected stable ES lines were treated with a limiting digestion of miccrococcal nuclease. DNA was recovered and tested by qPCR to determine the differential digestion conferred by nucleosome protection. The *Il12b* CGI insert has a clear nucleosome binding pattern, with high nucleosome occupancy over the first 140 bp and very low nucleosome occupancy over the last 130bp (Fig 3-1c). This pattern is recapitulated in multiple clones, and at the endogenous locus. Surprisingly, the nucleosome occupancy pattern is retained in each fragment of the *Il12b* CGI. The *Il12b* CGI 1-140bp fragment has high nucleosome occupancy across the insert. Conversely, the *Il12b* CGI 141-277 fragment has extremely low MNase protection and seems to be depleted of nucleosomes. Each fragment reflects its nucleosome occupancy in the full length insert. Notably, the depleted area overlaps the demethylation associated DNA sequence, strongly suggesting transcription factor binding is altering the local nucleosome landscape.

In addition to direct positioning, there is also evidence of positioning effects adjacent to the *Il12b* CGI insert. In *Il12b* CGI inserts that have high nucleosome occupancy, the immediately adjacent Insert Up region tends to be depleted. Conversely, when the insert site is depleted, Insert Up tends to have increased occupancy. A similar relationship can be found at the Insert Down region. Comparing the two *Il12b* CGI halves, 1-140 and 141-277, provides the most striking evidence for this effect: a change from nucleosome occupancy to depletion at the insert more than doubles the nucleosome density at the adjacent Insert Up and Down sites. Although we do not observe nucleosome depletion over the entire *Il12b* CGI insert, the notable depletion at the 207-270 fragment is evidence for strong positioning effects in relation to transcription factor binding.

Chromatin Properties of the Artificial Nucleosome Binding Sequence 601

601 is an artificial 150bp sequence that can perfectly position nucleosomes *in vitro* (Lowary and Widom, 1998). The positioning ability of 601 is conferred solely by sequence determinants. Interestingly for our study, the sequence of 601 is also CpG rich (Fig 3-3a). Several studies have attempted to describe the positioning affect of 601 at genomic locations *in vivo*, but the results remain decidedly mixed (Cole et al., 2012). None of these studies have yet considered the chromatin environment that may arise at the CpG rich 601 sequence.

Upon integration into genomic DNA, we find high DNA methylation at the 601 insert (Fig 3-3b, Fig 3-4d). The 70% average DNA methylation at 601 is similar to normal genomic background levels. This is in contrast to a repeat laden insert we designed, which acquired 90-100% methylation (Fig 3-4a). We conclude that 601 is not being specifically targeted for repression, but lacks the intrinsic ability to acquire low DNA methylation.

To determine if CpG content could alter the 601 chromatin environment, we created two 601 variant sequences. In one sequence we mutated all but three of the thirteen CpGs to GpC or GpG to create a Reduced CpG 601 construct (Fig 3-4b). We also introduced more CpGs to 601, targeting for mutation T-A dinucleotides that are though to help stabilize nucleosome binding via flexible base stacking (Vasudevan et al., 2010). This construct, which we named 601 Rigid, has higher CpG density and theoretically lower nucleosome binding ability (Fig 3-4c). Bisulfite sequencing of 601 variant BAC stable ES lines reveals that both modifications have DNA methylation similar to 601 (Fig 3-3c, Fig 3-8a). The retention of medium to high levels of methylation at 601 Rigid is reasonably surprising as it has 20 CpGs total, which is two more than the *Il12b* CGI insert. This extremely CpG dense 150bp insert is still unable to acquire low DNA methylation, suggesting high CpG density may not be sufficient to protect a region from DNA methylation.

The H3K4me3 levels at the 601 insert are near background levels (Fig 3-3d). The 601 Rigid construct has twice as much H3K4me3 as unmodified 601, but still only half of the moderate *Il12b* CGI H3K4me3 ChIP signal. Nucleosome occupancy at the 601 sequence is higher than seen in the *Il12b* CGI inserts. The occupancy at adjacent regions is also high suggesting that 601 may be increasing nucleosome density locally with heterogeneous positioning *in vivo*. Evidence that 601 may not be positioning nucleosomes perfectly like it does *in vitro* comes from the fact that the maximum protection score at the 601 insert never exceeds that seen at the reference Ebf1 nucleosome dense region or at the endogenous *Il12b* CGI 1-140bp (Fig 3-4d). Surprisingly, addition of CpGs did not destabilize the nucleosome binding of 601 Rigid; it has nucleosome occupancy comparable or higher than 601. Although the effect seen at 601 Reduced CpG is relatively minor, removal of CpGs did seem to reduce nucleosome density

at the insert. Regions adjacent to this insert also have high occupancy. Although 601 has a nucleosome positioning effect *in vivo*, it cannot provide information about the CpG island-nucleosome link as it does not acquire CpG island features.

Induction of CpG Island Features by a Transcription Factor

It has been shown that introduction of strong transcription factor binding sites can change local DNA methylation (Lienert et al., 2011; Macleod et al., 1994). We decided to use this system to study the putative transcription factor binding site in the *Il12b* CGI 207-270 fragment, and to attempt to introduce chromatin features to the 601 sequence. We cloned a direct fusion of 601 to the *Il12b* CGI fragment. Bisulfite sequencing in ES cell lines containing this BAC construct reveal that introduction of a nearby TF site is sufficient to spread demethylation of DNA entirely through the adjacent 150bp 601 sequence (Fig 3-5a, Fig 3-8b). CpG density was not required for the spread, as fusion of *Il12b* CGI 207-277 to the 601 Reduced CpG construct also resulted in low DNA methylation across the insert. Interestingly, although the *Il12b* CGI 207-277 fragment is sufficient to establish low DNA methylation, H3K4me3 remained much lower at both 207-277 fusion inserts than at the full length *Il12b* CGI insert (Fig 3-5b). This different is not explained by CpG density, as the 601 fusion construct has the same amount of CpGs over a shorter length as the *Il12b* CGI, yet it has only 40% of the H3K4me3 enrichment. It would appear that H3K4me3 levels are not solely determined by the availability of unmethylated CpGs. Additionally, the transcription factor that drives low methylation at the *Il12b* CGI does not directly recruit H3K4me3 at all; the 601 Reduced CpG insert and the 601 Reduced CpG + *Il12b* CGI 207-277 insert both had similar near-background levels of H3K4me3.

Next we tested the 601 + *Il12b* CGI 207-277 insert's nucleosome occupancy to determine how a switch to low DNA methylation altered the nucleosome profile over the 601 fragment. The nucleosome occupancy remains high over the 601 sequence, and remains low over the *Il12b* CGI 207-277 sequence (Fig 3-5c), suggesting integration of the individual profiles of the fused fragments (Fig 3-5d). This fusion also strongly resembles the histone profile at the full length *Il12b* CGI. This suggests that a major determinant of discreet histone positioning is the presence of a strongly depleted region which precisely orients adjacent histones. The nucleosome occupancy at 601 + *Il12b* CGI 207-277 is different enough from the full length *Il12b* CGI to suggest that the nucleosome binding strength of 601 has contributed to the overall profile. First, the nucleosome density seen at the *Il12b* CGI 207-277 portion of the fusion is much higher than seen there in previous experiments, even though the same BAC primer pairs were used. This putative transcription factor binding site sees a nearly 50% increase in MNase protection in this context. The extreme depletion at this site and at the adjacent Insert Down site seems to be mitigated by the presence of 601 upstream. Second, the overall nucleosome signal is much higher in the fusion construct than at the *Il12b* CGI insert, even though the profile shape is similar. The fusion sequence has a maximum nucleosome density that is 90% of the reference Ebf1 nucleosome, while the *Il12b* CGI insert sequence occupancy peaks at less than 50% of the reference nucleosome. These lines of evidence suggest that a primary positioning determinant is the nucleosome depletion caused by transcription factor binding, but the 601 sequence is also capable of affecting the resulting nucleosome density. As occupancy at the 601 portion of the fusion insert is not reduced by changing chromatin status, we conclude that low DNA methylation and nucleosome occupancy are not antagonistic at CpG islands.

Investigation of The *Il12b* CGI Transcription Factor and Its Role In DNA Methylation

Transcription factor binding site prediction at the *Il12b* CGI 207-277 fragment finds several protein binding motifs with high significance, the most interesting of which is the chromatin organizing zinc finger protein CTCF (Fig 3-6a). CTCF binding has been shown to position nucleosomes *in vivo,* and is also involved with maintenance of small low methylated windows genome wide (Fu et al., 2008; Stadler et al., 2011). CTCF has been shown to bind the *Il12b* CGI endogenous locus (Fig 3-6b). We've shown that the binding of CTCF is likely correlated with spreading of low DNA methylation and positioning adjacent nucleosomes. Chromatin immunoprecipitation at the *Il12b* CGI locus demonstrates that CTCF also appears to buffer the spread of H3K4me3 (Fig 3-6c). The moderate H3K4me3 signal at the *Il12b* CGI insert falls precipitously as it crosses the CTCF site, and the immediately adjacent Insert Down region has no signal.

To establish the importance of CTCF binding for the *Il12b* CGI, we created three constructs that remove the CTCF binding site; a 9bp deletion, a 15 bp deletion, and full deletion of the *Il12b* CGI 207-277 fragment, named *Il12b* CGI 1-206. In stable ES lines containing these BAC constructs, we find that ablation of the CTCF binding site removes the near-zero DNA methylation found previously (Fig 3-6d, Fig 3-9a). To our surprise, the *Il12b* CGI 1-206 fragment does not return to genomic background levels of methylation, instead remaining around 20-30% average DNA methylation. This result remained true across eight clones from two experimental replicates. It is unclear why the constructs with smaller deletions have higher DNA methylation than *Il12b* CGI 1-206, although it is worth noting that the *Il12b* CGI 15 bp deletion insert has lower average methylation compared to genomic background as well at around 40%. The prior *Il12b* CGI fragment analysis argues against the presense of another discreet

demethylating transcription factor binding site, as the sequence composing *Il12b* CGI 1-206 is within both of the highly methylated 1-140 and 114-206 fragments. The alternative suggestion is that a CpG dense fragment of a certain threshold length may be able to acquire low DNA methylation.

Ablation of the CTCF binding site does diminish H3K4me3 levels (Fig 3-6e). This is likely due to the higher DNA methylation caused by deletion of the CTCF binding site. Any H3K4me3 promoting sequence would be shared between the *Il12b* CGI CTCF deletions and full length *Il12b* CGI inserts, and we have shown previously that the CTCF binding site fragment does not increase H3K4me3 levels.

Evidence for DNA Methylation Protection Mediated by CpG Island Size in ES Cells

The low methylation seen at the *Il12b* CGI 1-206 fragment could not be explained by a deterministic transcription factor binding site, but instead may be triggered by an increase in CpG island size. The *Il12b* CGI 1-206 fragment is 66 bp longer than the heavily methylated 1-140 fragment. In order to determine whether the difference in sequence length was the critical factor, we cloned two *Il12b* CGI 1-140 fragments sequentially into the BAC insertion site, for a total length of 290bp. This tandem insert has the same CpG density and putative binding sites as the *in vivo* methylated *Il12b* CGI 1-140 fragment. Bisulfite sequencing of the tandem insert reveals that the increase in length was sufficient to incur a low DNA methylation state (Fig 6a, Fig 3-9b). The average methylation at the *Il12b* CGI 1-140 x2 insert across all CpGs and clones is 18%, which is lower than at the *Il12b* CGI 1-206 fragment (27%) but is higher than the near-zero methyaltion at the full length *Il12b* CGI insert (2%). The near-zero DNA methylation at the full length *Il12b* CGI sequence is likely contributed to by CTCF binding as previously discussed.

To remove the possibility that the low DNA methylation at the 1-140 tandem array occurred due to a unique property of the *Il12b* CGI sequence, we also cloned and tested sequential 601 sequences. The 601 tandem sequence, 601x2, contained a 20bp adapter and a total sequence length of 320bp. Bisulfite sequencing of the 601x2 insert, in stable ES cell clones, reveals that the DNA methylation status is much lower over the tandem array than at a single 601 insert (Fig 3-7b, Fig 3-9b). The average CpG methylation over 601x2 is 30% compared to 70% at a single 601 insert. Once again, a sufficiently sized CpG dense sequence was able to establish partial protection from DNA methylation through an unknown mechanism. To confirm that this effect was not caused by introduction of a transcription factor binding site in the linker DNA used to clone the arrays, we created a BAC insert composed of 601, the linker DNA, and the first part of the next 601 sequence. This sequence, named 601 + Adapter, retained high methylation in ES cells (Fig 3-7b).

The decrease in DNA methylation at the tandem inserts is complemented by an increase in H3K4me3 (Fig 3-7c). *Il12b* CGI 1-140 2x has 2.5 fold more H3K4me3 and 601x2 has 4 fold more H3K4me3 compared to single copy inserts. Mirroring our data with the *Il12b* CGI 207-277 fusion sequences, we find that the switch to low methylation at the 601x2 insert does not affect the nucleosome occupancy (Fig 3-7d). Increasing H3K4me3 over the 601x2 sequence has also not affected nucleosome occupancy. In fact, the nucleosome density over the tandem array insert is slightly increased compared to a single 601 copy. Another size threshold effect triggered by the sequential cloning insert was the one for targeted repression; 601x2 was the only sequence tested to acquire appreciable H3K27me3 (Fig 3-7e).

Discussion

Many studies of CpG islands occur at a large scale, leveraging genomics to study huge CpG stretches at promoters. Here we have described a small scale study that allows us to interrogate some of the inner workings of the CpG island, and how its properties relate to each other.

We find that even at a small CpG island, transcription factor binding seems to be a key event. The manner in which a small binding site can establish low methylation and histone depletion, which can be spread into the surrounding chromatin, could easily be duplicated at CpG islands genome wide. Indeed, the protein CTCF does bind at many sites with low DNA methylation across the mouse genome (Stadler et al., 2011). It has also been shown that CTCF binding is correlated with 5-hydroxymethylcytosine, a marker for active demethylation (Feldmann et al., 2013). Here we show that CTCF binding may spread low DNA methylation into regions of high CpG density, which could be an important contributor to the methylation state at CpG islands genome wide.The evidence that CTCF has a functional role in demethylating DNA surrounding its binding sites is so far only correlative, and no compelling mechanisms have been proposed.

Another interesting facet of CTCF binding is the effect it has on nucleosome positioning. If CTCF has a functional role binding to CpG islands and promoting local demethylation, it may be possible that discreet histone positioning is also important to CpG island regulation. The *Il12b* CGI is not a promoter CpG Island; the *Il12b* promoter is methylated in ES cells. The CTCF lies between the majority of the CpG island and the promoter. It is possible that the strong positioning effect of the CTCF site in the *Il12b* CpG island blocks aberrant expression at the promoter by controlling nucleosome density there. The CTCF binding site also was shown to

buffer the H3K4me3 arising at the *Il12b* CGI from spreading towards the inactive promoter. This role is supported by the observation that the CTCF site does not increase H3K4me3 enrichment in most contexts, only affecting local DNA methylation. The CTCF site is approximately 880 bp away from the transcription start site, which places it 3-4 nucleosomes away from the core promoter. The *Il12b* promoter has a crucial nucleosome in macrophages (Ramirez-Carrozzi et al., 2006), which may be positioned by the upstream CTCF site.

While CTCF binding is a simple mechanism to understand and test, the effect of underlying nucleotide content has remained difficult to elucidate. Here we have described the interaction between 601, the best *in vitro* positioning sequence, and CpG island chromatin features. Although we find that 601 does not have precise positioning *in vivo*, it still effectively increased the local histone density regardless of the level of DNA methylation. Current opinions on histone positioning, including what has been learned from 601, support high GC content as a strong determinant of intrinsic histone positioning in yeast (Kaplan et al., 2010), which is difficult to reconcile with the nucleosome depletion seen at mammalian CpG island promoters. We found evidence that at non-promoter regions, inceasing GC content seemed to increase the nucleosome occupancy at our constructs and vice versa in murine cells. However we also found that the 601 nucleosome affinity was dominated by the nearby presence of the CTCF binding site. Transcription factor binding is a common occurance at promoters, and may be the dominant positioning determinant there. One caveat is that our studies did not consider sequences large enough to bind multiple nucleosomes. Large stretches of CpG rich DNA may behave differently once the length exceeds two bound nucleosomes.

Curiously, increasing the size of the CpG island via sequential cloning was the only modification we found that increased H3K4me3 a great deal. H3K4me3 levels were refractory to

transcription factor binding, increases in CpG density, and changes in nucleosome occupancy, but increased 2-4 fold as the size of the CpG island doubled. Likewise, the 601-*Il12b* CGI 207-277 fusion had lower H3K4me3 then full length *Il12b* CGI despite similar features. One key difference is that the fusion is 57bp shorter. The limiting factor altered by size may be nucleosome substrate for recruited histone methyltransferases to modify. A longer sequence has unmethylated CpGs which overlap more nucleosomes. Unlike a model based solely on discreet binding sites in all CpG islands, this mechanism provides a possible rationale for the large CpG island sizes found genome wide.

Perhaps most strikingly, our study describes an alternate method of CpG island protection from DNA methylation that is based on length. Three lines of evidence suggest are findings are not due to aberant introduction of a new transcription factor binding site: first, we have tested and ruled out the linker DNA, second, the sequences used in the tandem inserts were shown to have no demethylation activity at all alone, and third, the DNA methylation level is low but different than that seen at the inserts with a strong transcription factor binding site. We also were able to show that increasing CpG density cannot trigger the same effect; the extremely CpG dense 601 Rigid remained methylated.

Size threshold based protection from DNA methylation does not seem to act through nucleosome remodeling, as nucleosome occupancy remains high in the 601x2 construct. Additionally the protection from DNA methylation is weak and seems to have specific sequence determinants beside CpG content; demonstrated by the slightly lower methylation at *Il12b* CGI 1-140 x2 insert despite a smaller size and fewer CpGs than 601x2. A likely explanation may be binding of non-specific transcription factors which favor CpGs in their binding sites and have low binding affinity. The size threshold may trigger protection from DNA methylation by

increasing the overall chance of weakly binding a general factor, or multiple factors. It is also not possible to entirely rule out the role of nucleosomes in this process, indeed it is intriguing that the size threshold to trigger low methylation seems to be between 140 and 200bp. An obvious known threshold around that size is the length of DNA needed to fully wrap around a nucleosome; 150bp.

What role do all of these features play together? We speculate that the size related protection from methylation may help initiate low methylation which can assist in initial binding of proteins like CTCF or maintain a similar CpG island if a strong transcription factor is repressed. The size of a CpG island also increases its H3K4me3 content. The *Il12b* CGI may use two DNA methylation strategies and high H3K4me3 to retain euchromatin throughout differentiation. This could facilitate the binding of cell type specific factors (for instance *Il12b* CGI contains an Elk-1 site) while maintaining silencing at the promoter. We have demonstrated in this study that the chromatin properties at CpG Islands can be modulated and fine tuned to an impressive degree by controlling the sequence length and binding sites. Doubtlessly the cell has evolved to capitalize on this flexibility to help modulate gene expression. A combination of large scale genomic studies and continued small scale manipulative experiments will help yield understanding of the logic of CpG island regulation.

Materials and Methods

Cell culture and reagents

The R1 murine ES line was grown in Knockout DMEM supplemented with 15% fetal bovine serum (Omega), 0.1 mM nonessential amino acids, 2 mM L-glutamine, 1% penicillin/streptomycin, 0.05 mM β-mecaptoenthanol, and 1000 U/ml LIF (ESGRO, Millipore).

All culture products were purchased from Gibco unless otherwise noted. ES cells were maintained in gelatin (Stem Cell Technologies) coated Petri dishes and on a layer of mouse embryonic fibroblasts mitotically inactivated with mytomycin-C. ES cells were removed from plates using Trypsin-EDTA (Stem Cell Technologies) for 5 minutes which was then neutralized by FBS containing media.

BAC Modification and Preparation

The human gene desert RP11-722D BAC was purchased from CHORI-BACPAC. Insertion of exogenous sequence into the BAC was done according to a protocol adapted from (Gong and Yang, 2005). BACs were electroporated into SW102 RecA expressing bacteria and selected for targeted recombination of GalK and replacement of GalK by minimal galactose media or deoxygalactose respectively (Warming et al., 2005). For stable ES cell transduction, a PGK-Neomycin expressing cassette was introduced into the BAC as described in (Wang, 2001). Successful recombineering was confirmed by restriction enzyme fingerprinting and sequencing of the insert region.

To prepare for ES cell transduction, BAC DNA was isolated using the Large Construct Kit (Qiagen) and linearized with the restriction enzyme PI-SceI. Pre-methylation of BACs was done by overnight incubation with *Sss*I methylase and SAM. BAC DNA was then phenol chloroform extracted and resuspended in 500uL PBS for electroporation. BAC integrity was verified on a large pulse field gel (BIO-Rad CHEF Mapper XA).

Generation of Stable ES Cell BAC lines

ES cells were grown to confluency in a 10cm plate prior to transduction by 5-20ug of BAC DNA by electroporation at 0.27kV 500uFd. After a short recovery, ES cells were replated 1:2. Selection for BAC integration was done using the antibiotic G418/Neomycin at 255ug/ul for approximately ten days. At this point single colonies were picked and outgrown into stable clones, maintained in G418. Genomic DNA was isolated from stable ES clones with the DNeasy kit (Qiagen). Integration of BAC DNA was confirmed by genotyping PCR.

Bisulfite Sequencing

Bisulfite treatment of 2.5ug of genomic DNA was performed overnight at 55 C, following denaturation by 5ul of 3M NaOH. The bisulfite-treated DNA was desalted using the PCR Purification kit (Qiagen), then was neutralized with 5m ammonium acetate and precipitated with 2mg yeast tRNA. Bisulfite-treated DNA was resuspended in 50ul TE.

Sequence-specific PCR of the bisulfite-treated DNA was performed using primers specific to BAC regions. The PCR fragments were cloned into the pCRII vector (Invitrogen, K2070-20) and transformed into DH5α *E. coli* cells. Miniprep plasmid DNA was sequenced using M13 reverse primers.

Chromatin Immunoprecipitation (ChIP)

Approximately 30 million ES cells were trypsinized, washed, and treated with 1% formaldehyde for 10 minutes. After neutralization with Glycine, cells were washed with PBS and treated with cell lysis buffer(5mM PIPES, 85mM KCl, 0.5% NP-40) for 10 minutes, and then nuclei lysis buffer (50nM Tris HCl, 10mM EDTA, 1% SDS) for 10 minutes. The nuclei were supplemented with protease inhibitors (α-Complete, Roche) and sonicated in a Diagenode Biorupter Twin sonicator for 15 minutes with 30 second cycles. Chromatin was frozen at -70 C

until use. After thaw and removal of SDS, 100ug of chromatin was incubated overnight at 4 C with 250 ul of ChIP dilution buffer and 5 ug per antibody; we used H3K4me3 (Millipore, 07-473) and H3K27me3 (Active Motif, 39155). Immune chromatin complexes were recovered by binding to Protein A Dynabeads (Invitrogen, 100-02D) for 20 minutes at 4 C and 20 minutes at room temperature, and then isolated and washed with a magnet. IPed chromatin was released from the Protein A beads by elution with $NaCHO_3$ 1% SDS buffer, and crosslinking was reversed by incubation at 65 C overnight. DNA was purified using the PCR purification kit (Qiagen).

Quantity of immunoprecipitated DNA was measured by qPCR on an iCycler (BioRad). Three or four sets of primers were designed specific to each BAC insert, and tested for PCR efficiency and a consistent melt curve. Each run included endogenous control primers and BAC non-insert primer sets. The amount of IPed DNA for each primer set was calculated relative to a 5% input chromatin control sample. To control for variable BAC integrants, the % input for all BAC regions were normalized to the % input at a Downstream BAC CpG island with consistent enrichment.

MNase Protection Assay

Approximately 10 million ES cells were trypsinized, washed, and treated with cell lysis buffer (5mM PIPES, 85mM KCl, 0.5% NP-40) for ten minutes. Nuclei were resuspended in 750 ul MNase digestion buffer (50mM TrisHCl, 1mM $CaCl_2$, 0.2% Triton X-100) , divided into 150ul aliquotes and preheated to 37 C for two minutes. Miccrococcal nuclease (Worthington, Ls 4797) was added at 1 Unit/ ul and the digestion was immediately incubated at 37 C for either 5

minutes or 30 minutes. Digestion was stopped with stop solution supplemented with Protinase K and incubated at 37 overnight. DNA was recovered by phenol chloroform extraction.

Cleavage was quantified by qPCR. Primer sets specific to the BAC insert and control regions were designed to be approximately 100-120bp. MNase protection was calculated from the threshold cycle difference between MNase treated samples and uncut DNA. All MNase delta C(t)s were normalized to the control nucleosome high and low regions near the Ebf1 CpG island, where 100% is the same occupancy as the Ebf1 nucleosome and 0% is nucleosome level at the depleted Ebf1 linker region downstream.

Transcription Factor Binding Datasets

Transcription factor binding sites were discovered using the Jaspar database on PSCAN (Zambelli et al., 2009). The *Il12b* CGI and CTCF ChIP-sequencing data was visualized using UCSC Genome Browser (Kent et al., 2002). CTCF ChIP sequencing tracks are from: ES Cells (Stadler et al., 2011) and Frontal Cortex (Bing Ren's laboratory, ENCODE/LICR).

Figure Legends

Figure 3-1 – Examination of Chromatin Properties by Fragmentation of a CpG Island

(A) DNA methylation status of *Il12b* CGI and fragments inserted into a stably integrated BAC in ES cells, determined by bisulfite sequencing. Each CpG had at least 10 fold coverage for at least 3 separate clones. Color in each ball and stick represents the average DNA methylation at a specific CpG, indicated by colors described in legend. Each fragment is shown vertically aligned to the full *Il12b* CGI. (B) Chromatin immunoprecipitation for H3K4me3 at each BAC insert. A downstream CGI on the BAC with sequence features comparable to the *Il12b* CGI is used as a positive control. Relative % input is determined by normalization to the BAC Downstream CGI % input. Color indicates DNA methylation status. Error bars represent standard error. (C) MNase protection qPCR assay at the *Il12b* CGI and fragments. Background is 4k upstream, Insert Up and Down are directly adjacent, and Insert 1 and 2 sites are within BAC insert. Shown is protection from MNase, normalized to a genomic nucleosome control. Higher values indicate higher nucleosome occupancy.

Figure 3-2 – DNA Methylation at the *Il12b* CpG Island in a Gene Desert BAC (A) Genomic location of the 136kb human gene desert BAC used in this study, from UCSC Genome Browser. The insertion site is marked, nearly 1Mb away from the closest promoter element. (B) Bisulfite sequencing data for the *Il12b* CGI BAC insert, and the fragments 1-140, 141-270, and 207-270, in stable ES cell lines. CpG positions are indicated to the left, with % methylated CpGs and ratio of methylated CpGs to total CpGs given across the table, for each condition at top. The first column for each BAC construct contains confirmation of the *in vitro* BAC pre-methylation with SssI. Colors indicate methylation status, see key. (C) Chromatin immunoprecipitation for

H3K4me3, shown is raw % input for the endogenous *Il12b* CGI, active house-keeping genes, and the inactive *Il12b* promoter in ES cells. (D) H3K4me3 ChIP at the endogenous *Il12b* CGI and *Il12b* CGI BAC insert, over 2 clones and three replicates. Standard error bars are shown.

Figure 3-3 – Chromatin Properties of the 601 Positioning Sequence (A) The 601 sequence with CpGs highlighted (B) DNA methylation status at the 601 sequence and at (C) 601 variants inserted into a stably integrated BAC in ES cells, determined by bisulfite sequencing, as in Fig 3-1. Methylation status is indicated by color, described in legend. (D) Chromatin immunoprecipitation for H3K4me3 at each BAC insert, with the addition of the *Il12b* CGI insert from Fig 3-1b for comparison. (E) MNase protection qPCR assay at the 601 sequence and variant BAC insertions. Shown is protection from MNase, normalized to a genomic nucleosome control. Higher values indicate higher nucleosome occupancy.

Figure 3-4– Repetitive DNA Insertion and 601 Bisulfite Sequencing

(A) DNA methylation status at the Lox 5x Gal DBS sequence inserted into a stably integrated BAC in ES cells, determined by bisulfite sequencing, as in Fig 3-1. Methylation status is indicated by color, described in legend. This short sequence contains floxed Gal (a non-mouse yeast activator) repeat binding sites. (B) Design of the 601 Reduced CpG insert sequence. CpGs are highlighted in grey, changes are shown in lower case. (C) Design of the 601 Rigid insert sequence. Yellow highlights the flexible regions in the 601-nucleosome structure. Blue highlights the target “flexible” nucleotides which were mutated. Substituted bases are in lower cases. (D) Bisulfite sequencing data for 601 BAC insert in stable ES lines. 601 CpG position is shown on left, methylation is indicated by color, in key. (E) Raw qPCR data from an MNase

protection assay for a 601 integrated ES clone with the best insert protection score. Shown is the C(t) difference between cut and uncut DNA for various primers, and three different primers that cover the BAC insert.

Figure 3-5 – The Effect of Transcription Factor Binding on Local CpG Island Chromatin (A) DNA methylation status at the 601 and 601 Reduced CpG sequences fused to *Il12b* CGI 207-277, inserted into a stably integrated BAC in ES cells, as determined by bisulfite sequencing. As in Fig 3-1a, methylation status is indicated by color, described in legend. (B) Chromatin immunoprecipitation for H3K4me3 at each BAC insert, with the addition of *Il12b* CGI from Fig 3-1b for comparison. (C) MNase protection qPCR assay at the 601+ *Il12b*CGI 207-277 BAC insert. Shown is protection from MNase, normalized to a genomic nucleosome control. Higher values indicate higher nucleosome occupancy. (D) Overlay of the 601-*Il12b* CGI 207-277 MNase diagram with the MNase diagrams from the fusion pieces alone.

Figure 3-6 – A Putative CTCF Site in the *Il12b* CGI is Sufficient but Not Necessary for Low DNA Methylation (A) Transcription factor binding site analysis of the *Il12b* CGI 207-277 fragment. Highest hits are shown below the sequence they bind (B) CTCF binding at the *Il12b* CpG island in ES cells and Frontal cortex cells is shown in UCSC genome browser, 1 kb upstream of the *Il12b* gene. The frontal cortex CpG methylome data is included above the CTCF ChIP Seq tracks. (C) H3K4me3 chromatin immunoprecipitation for primers across the *Il12b* CGI BAC insert (D) DNA methylation status at the *Il12b* CTCF site deletion mutants, inserted into a stably integrated BAC in ES cells, determined by bisulfite sequencing as in Fig 3-1. The 15bp and 9bp deletions both overlap the CTCF site. Methylation status is indicated by color, described

in legend. (E) Chromatin immunoprecipitation for H3K4me3 at each *Il12b* CGI CTCF deletion BAC insert, with the addition of *Il12b* CGI from Fig 3-1b for comparison

Figure 3-7 – Sufficient CpG Island Size Can Trigger a Low DNA Methylation State

(A) Comparison of DNA methylation status at the *Il12b* CGI 1-140 fragment from Fig 3-1a and at the *Il12b* CGI 1-140 x2 construct, inserted into a stably integrated BAC in ES cells, determined by bisulfite sequencing as in Fig 3-1. Methylation status is indicated by color, described in legend. (B) Comparison of DNA methylation status at the 601 sequence, at two 601 sequences joined together (601x2) and at 601 with the adapter sequence only, inserted into a stably integrated BAC in ES cells. (B) Chromatin immunoprecipitation for H3K4me3 at each BAC insert, with the addition of *Il12b* CGI from Fig 3-1b for comparison. (C) MNase protection qPCR assay at the 601x2 BAC insert, with 601 for comparison from Fig 3-3e. Shown is protection from MNase, normalized to a genomic nucleosome control. Higher values indicate higher nucleosome occupancy. (E) Chromatin immunoprecipitation for H3K27me3 at several BAC inserts. % input is normalized to the bivalent HoxA7 promoter.

Figure 3-8 – Bisulfite Sequencing for 601 Variants and Fusion Inserts

(A) Bisulfite sequencing for 601 Variant BACs in stable ES lines; 601 Reduced CpG and 601 Rigid. CpG positions are indicated to the left. The first column for each BAC construct contains confirmation of the BAC premethyaltion. Colors indicate methylation status, see key. (B) Bisulfite sequencing data for the 601 + *Il12b* CGI 207-277 BAC insert or 601 Reduced CpG + *Il12b* CGI 207-277 BAC insert, in stable ES cell lines.

Figure 3-9- Bisulfite Sequencing for CTCF Deletions and Tandem Arrays

(A) Bisulfite sequencing of Il12b CGI 15bp Deletion and Il12b CGI 1-206 BACs stably transfected into ES cells. CpG positions are indicated to the left. Colors indicate methylation status, see key. (B) Bisulfite sequencing of sequential array BAC inserts *Il12b* CGI 1-140 x2 and 601 x2 in ES cells.

Figure 3-1 – Examination of Chromatin Properties by Fragmentation of a CpG island

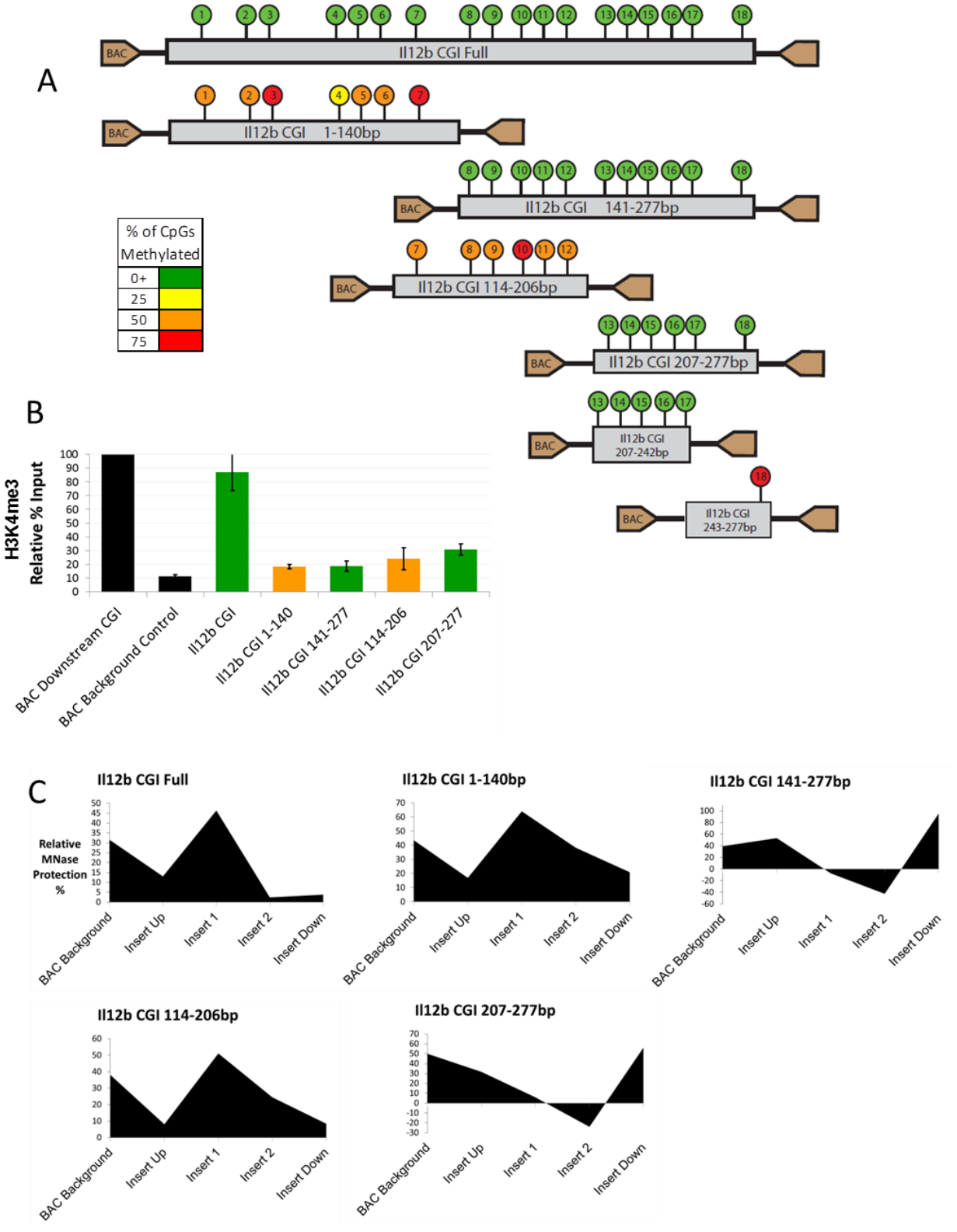
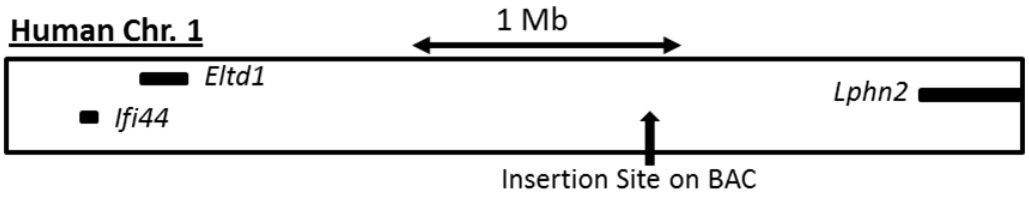


Figure 3-2 – DNA Methylation at the *Il12b* CpG Island in a Gene Desert BAC

A



B

ESC clones with Pre-Methylated Il12b CGI BAC

Il12b CpG Island Insert	Clone 1 %	Clone 1 Ratio	Clone 2 %	Clone 2 Ratio	Clone 3 %	Clone 3 Ratio
-1054	9	1/11	0	0/10	10	1/10
-1041	9	1/11	10	1/10	0	0/10
-1037	0	0/11	0	0/10	0	0/10
-990	0	0/11	0	0/10	0	0/10
-957	9	1/11	0	0/10	0	0/10
-938	0	0/11	0	0/10	0	0/10
-929	9	1/11	0	0/10	10	1/10
-927	0	0/11	0	0/10	0	0/10
-915	0	0/11	0	0/10	10	1/10
-910	0	0/11	0	0/10	0	0/10
-908	0	0/11	0	0/10	0	0/10
-898	0	0/11	0	0/10	0	0/10
-895	0	0/11	0	0/10	0	0/10
-866	0	0/11	0	0/10	0	0/10
-846	0	0/11	0	0/10	0	0/10
-836	0	0/11	0	0/10	0	0/10
-833	0	0/11	0	0/10	0	0/10
-804	0	0/11	10	1/10	0	0/10

ESC clones with Pre-Methylated Il12b CGI 1-140 BAC

Pre-Meth BAC %	Pre-Meth BAC Ratio	Clone 1 %	Clone 1 Ratio	Clone 2 %	Clone 2 Ratio	Clone 3 %	Clone 3 Ratio
100	7/7	69	11/16	69	11/16	80	12/15
86	6/7	75	12/16	75	12/16	63	10/16
86	6/7	100	16/16	69	11/16	69	11/16
71	5/7	63	10/16	19	3/16	50	8/16
71	5/7	81	13/16	50	8/16	63	10/16
71	5/7	63	10/16	75	12/16	63	10/16
71	5/7	94	15/16	69	11/16	88	14/16

% of Methylated CpG: 0, 20, 40, 60, 80+

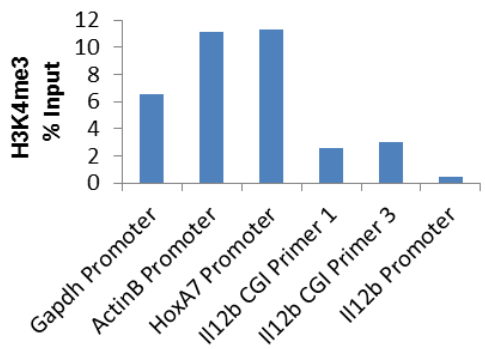
ESC clones with Pre-Methylated Il12b CGI 141-270 BAC

Pre-Meth BAC %	Pre-Meth BAC Ratio	Clone 1 %	Clone 1 Ratio	Clone 2 %	Clone 2 Ratio	Clone 3 %	Clone 3 Ratio	Clone 4 %	Clone 4 Ratio
88	7/8	10	1/10	0	0/10	0	0/10	0	0/10
100	8/8	20	2/10	10	1/10	0	0/10	0	0/10
100	8/8	0	0/10	0	0/10	10	1/10	0	0/10
100	8/8	0	0/10	0	0/10	0	0/10	0	0/10
100	8/8	0	0/10	0	0/10	0	0/10	0	0/10
75	6/8	0	0/10	0	0/10	0	0/10	0	0/10
100	8/8	0	0/10	0	0/10	0	0/10	0	0/10
100	8/8	0	0/10	0	0/10	0	0/10	0	0/10
88	7/8	0	0/10	0	0/10	0	0/10	0	0/10
100	8/8	0	0/10	0	0/10	0	0/10	0	0/10
100	8/8	0	0/10	0	0/10	0	0/10	0	0/10

ESC clones with Pre-Methylated Il12b CGI 207-270 BAC

Pre-Meth BAC %	Pre-Meth BAC Ratio	Clone 1 %	Clone 1 Ratio	Clone 2 %	Clone 2 Ratio	Clone 3 %	Clone 3 Ratio
100	7/7	13	2/16	0	0/14	0	0/16
100	7/7	0	0/16	0	0/14	6	1/16
100	7/7	0	0/15	7	1/14	0	0/16
71	5/7	0	0/16	0	0/14	0	0/16
86	6/7	0	0/16	7	1/14	6	1/16
100	7/7	19	3/16	7	1/14	19	3/16

C



D

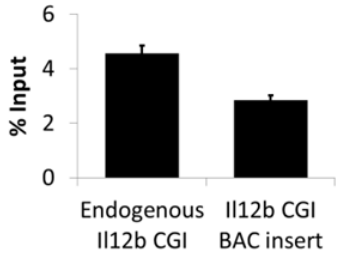


Figure 3-3 – Chromatin Properties of the 601 Positioning Sequence

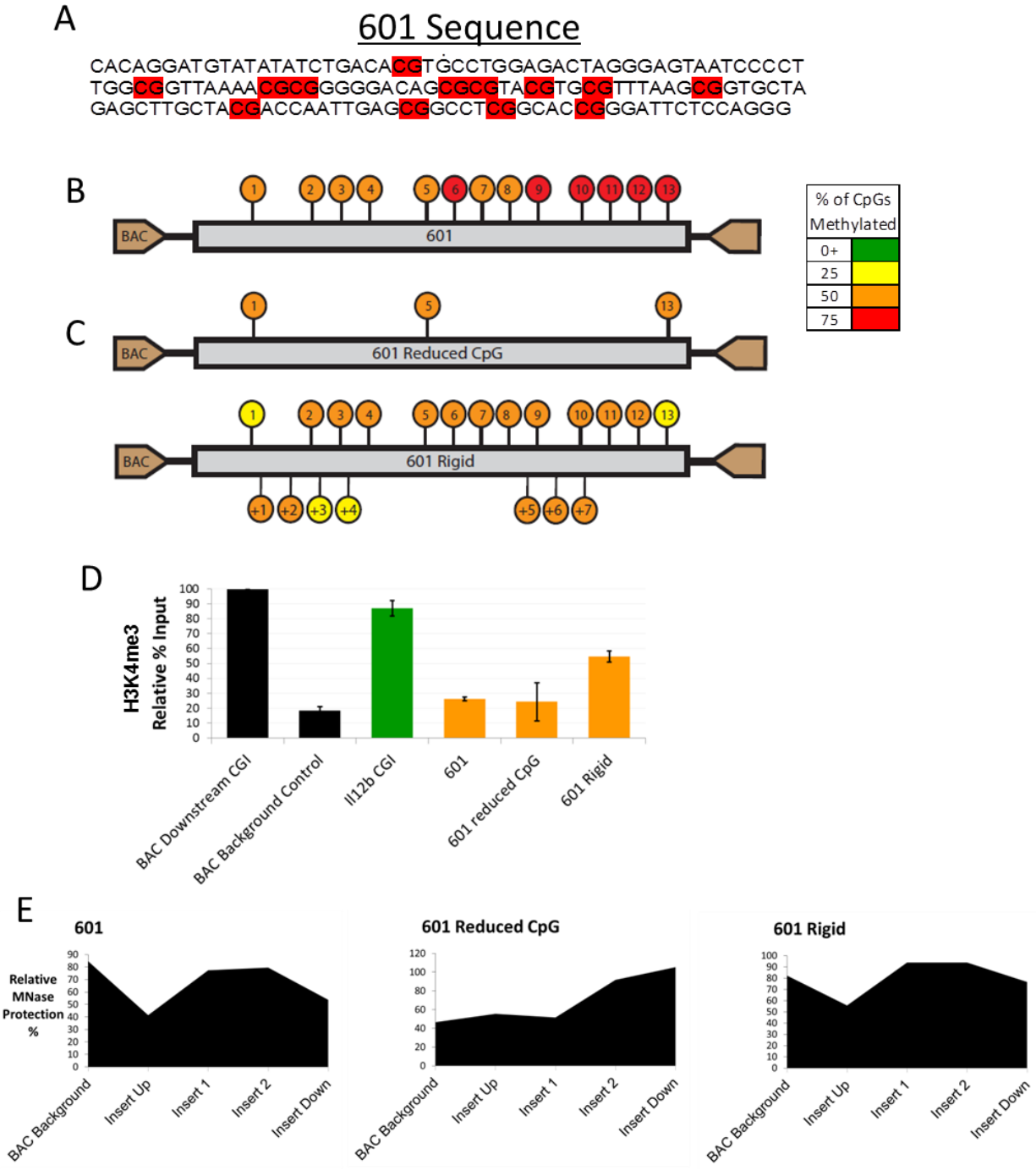
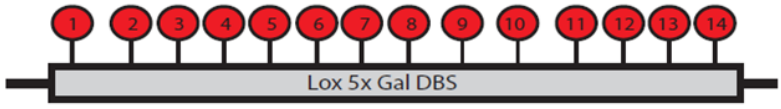
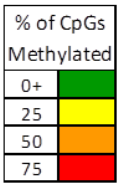


Figure 3-4 – Repetitive DNA Insertion and 601 Bisulfite Sequencing

A





B

601 Reduced CpG

CACAGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGgGGTT
AAAAgGaGGGGGACAGCGgcTAgcTGgcTTTAAGgGGTGCTAGAGCTTGCTAgcACCAA
TTGAGgGGCCTggGCACCGGGATTCTCCAGGG

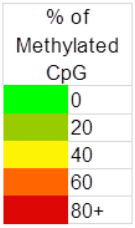
C

601 Rigid

CACAGGATGTATATATCTGcCACGTGCCgcGAGACgcGGGAGTAATCCCCTTGGCGGTg
cAAACGCGGGGGACAGCGCGgcCGTGCGTTgcAGCGGTGCgcGAGCTTGCgcCGACCAA
TgcAGCGGCCTgcGCACCGGGgcTCTCCAGGG

D

		ESC clones with 601 BAC											
		Pre-Meth BAC		Clone 1		Clone 2		Clone 3		Pre-Me clone 1		Pre-Me clone 2		Pre-Me clone 3	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
601 Insert	1	100	7/7	43	3/7	63	5/8	100	5/5	25	2/8	29	2/7	83	5/6
	2	100	7/7	43	3/7	75	6/8	100	5/5	75	6/8	43	3/7	67	4/6
	3	100	7/7	29	2/7	63	5/8	100	5/5	38	3/8	43	3/7	100	6/6
	4	100	7/7	29	2/7	63	5/8	80	4/5	38	3/8	43	3/7	83	5/6
	5	100	7/7	29	2/7	63	5/8	100	5/5	75	6/8	80	4/5	100	6/6
	6	100	7/7	71	5/7	100	8/8	100	5/5	75	6/8	80	4/5	100	6/6
	7	100	7/7	14	1/7	100	8/8	60	3/5	25	2/8	40	2/5	100	6/6
	8	86	6/7	43	3/7	63	5/8	80	4/5	38	3/8	40	2/5	100	6/6
	9	100	7/7	29	2/7	100	8/8	100	5/5	88	7/8	40	2/5	100	6/6
	10	100	7/7	57	4/7	88	7/8	100	5/5	100	8/8	40	2/5	83	5/6
	11	86	6/7	71	5/7	100	8/8	80	4/5	63	5/8	60	3/5	100	6/6
	12	100	7/7	86	6/7	100	8/8	80	4/5	75	6/8	60	3/5	67	4/6
	13	86	6/7	43	3/7	100	8/8	100	5/5	75	6/8	80	4/5	83	5/6



E

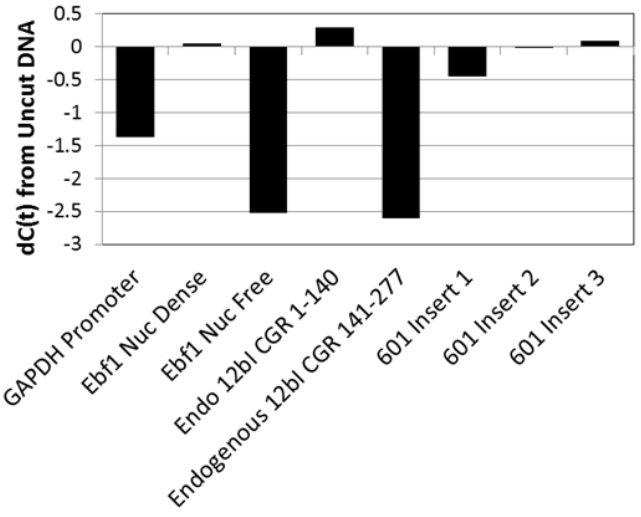


Figure 3-5 – The Effect of Transcription Factor Binding on Local CpG Island Chromatin

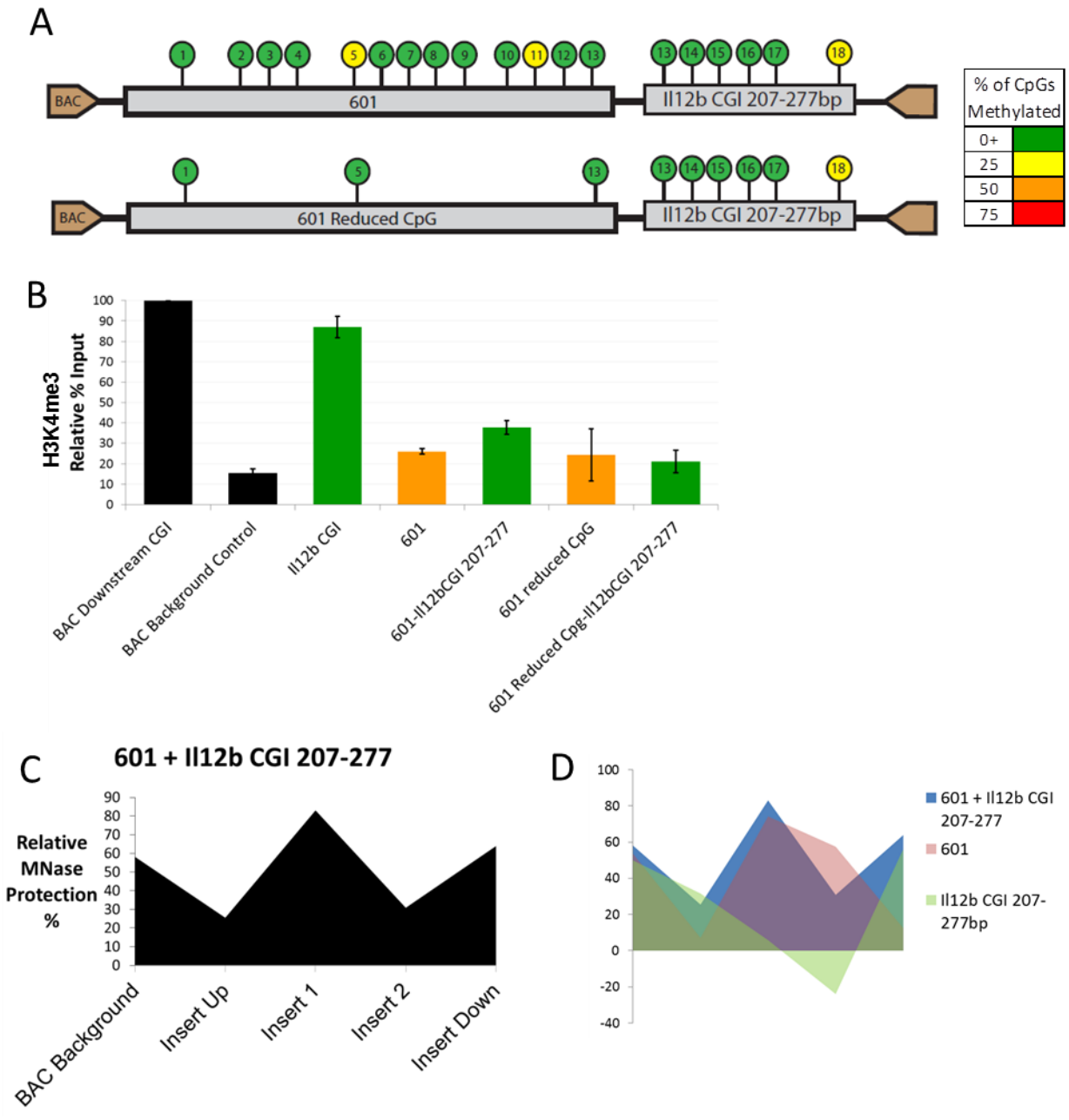


Figure 3-6 – A Putative CTCF Site in the *Il12b* CGI Is Sufficient But Not Necessary for Low DNA Methylation

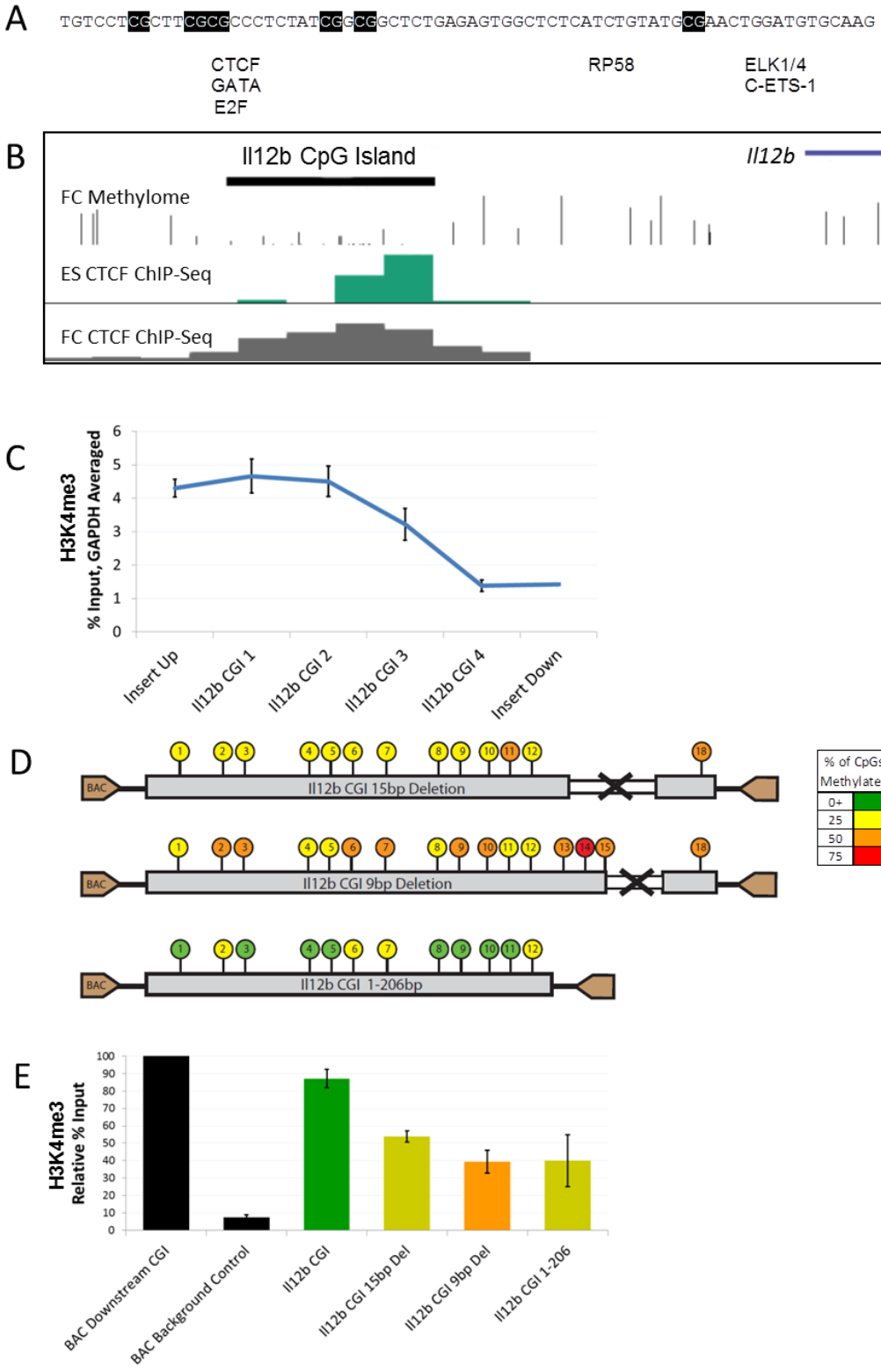


Figure 3-7 – Sufficient CpG Island Size Can Trigger a Low DNA Methylation State

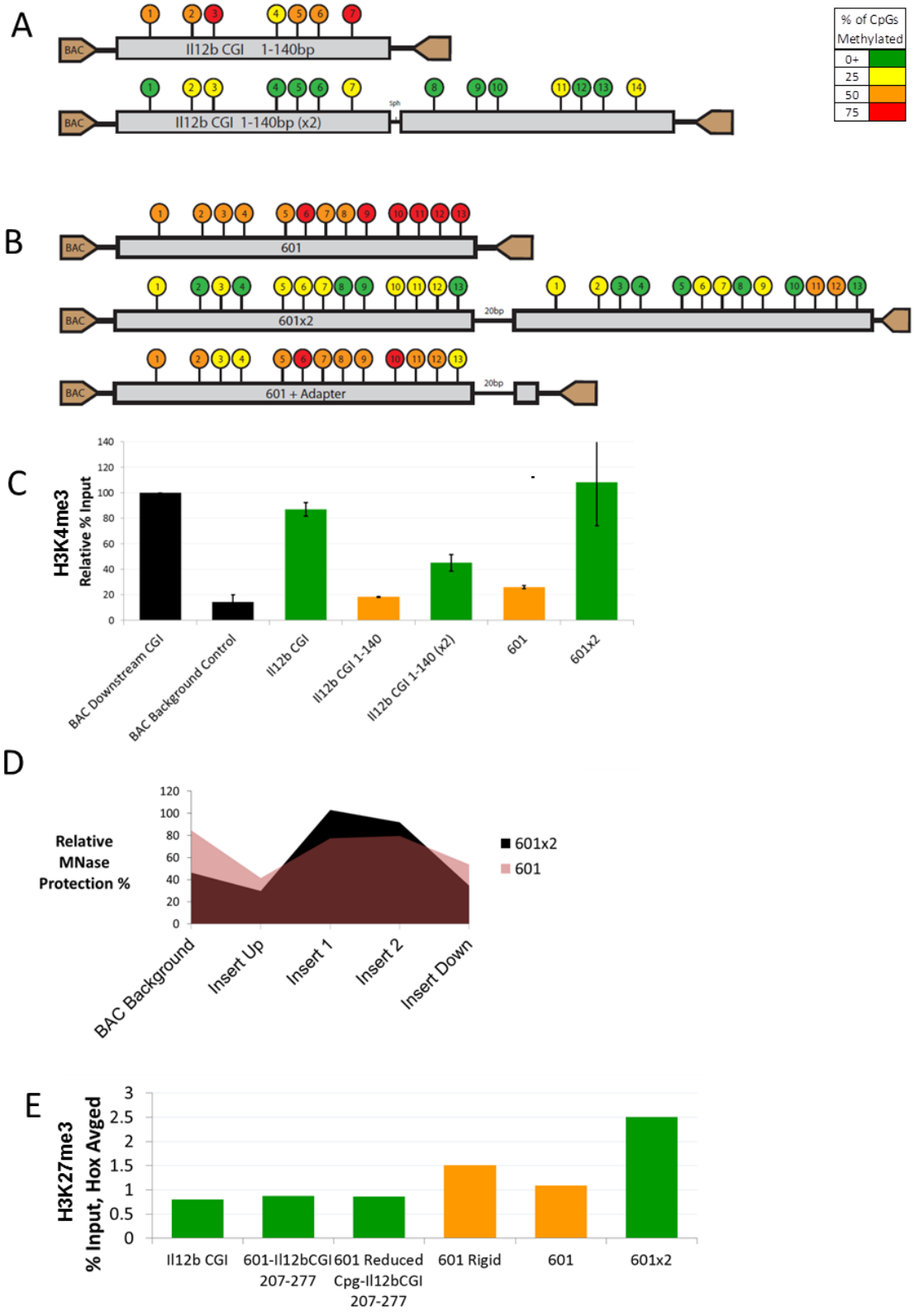
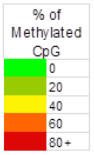


Figure 3-8 – Bisulfite Sequencing for 601 Variants and Fusion Inserts

A

		ESC clones with 601 Reduced CpG BAC																	
		Clone 1		Clone 2		Clone 3		Clone 4		Clone 5		Clone 6		Pre-Me clone 1		Pre-Me clone 2		Pre-Me clone 3	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
601 Reduced CpG Insert	1	89	8/9	100	9/9	60	6/10	50	4/8	100	9/9	100	9/9	33	4/12	42	5/12	50	5/10
	2	67	6/9	78	7/9	60	6/10	63	5/8	90	9/10	67	6/9	17	2/12	67	8/12	40	4/10
	3	70	7/10	78	7/9	80	8/10	63	5/8	90	9/10	89	8/9	42	5/12	58	7/12	60	6/10

		ESC clones with Pre-Methylated 601 Rigid BAC									
		Pre-Meth BAC		Clone 1		Clone 2		Clone 3		Clone 4	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
601 Rigid Insert	1	100	8/8	8	1/13	50	7/14	73	8/11	8	1/13
	2	100	8/8	31	4/13	86	12/14	91	10/11	8	1/13
	3	100	8/8	54	7/13	79	11/14	64	7/11	23	3/13
	4	100	8/8	38	5/13	79	11/14	82	9/11	15	2/13
	5	100	8/8	15	2/13	86	12/14	73	8/11	23	3/13
	6	88	7/8	46	6/13	79	11/14	73	8/11	15	2/13
	7	100	8/8	23	3/13	57	8/14	100	11/11	15	2/13
	8	100	8/8	54	7/13	86	12/14	100	11/11	31	4/13
	9	100	8/8	54	7/13	79	11/14	91	10/11	31	4/13
	10	100	8/8	46	6/13	86	12/14	91	10/11	46	6/13
	11	100	8/8	54	7/13	79	11/14	73	8/11	15	2/13
	12	100	8/8	38	5/13	86	12/14	91	10/11	46	6/13
	13	88	7/8	54	7/13	79	11/14	91	10/11	38	5/13
	14	100	8/8	54	7/13	92	12/13	91	10/11	46	6/13
	15	100	8/8	38	5/13	69	9/13	91	10/11	54	7/13
	16	100	8/8	54	7/13	85	11/13	91	10/11	54	7/13
	17	100	8/8	23	3/13	77	10/13	91	10/11	38	5/13
	18	100	8/8	69	9/13	85	11/13	82	9/11	46	6/13
	19	100	8/8	54	7/13	85	11/13	91	10/11	46	6/13
	20	88	7/8	15	2/13	46	6/13	82	9/11	38	5/13



B

		ESC clones with Pre-Methylated 601 + Il12b CGI 207-277 BAC									
		Pre-Meth BAC		Clone 1		Clone 2		Clone 3		Clone 4	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
601	1	83	5/6	0	0/12	17	2/12	8	1/12	0	0/12
	2	83	5/6	0	0/12	42	5/12	8	1/12	8	1/12
	3	100	6/6	8	1/12	0	0/12	0	0/12	0	0/12
	4	80	4/5	8	1/12	17	2/12	25	3/12	17	2/12
	5	100	5/5	25	3/12	33	4/12	50	6/12	33	4/12
	6	100	6/6	17	2/12	8	1/12	25	3/12	25	3/12
	7	100	4/4	25	3/12	17	2/12	25	3/12	25	3/12
	8	100	5/5	25	3/12	0	0/12	33	4/12	0	0/12
	9	100	5/5	17	2/12	0	0/12	33	4/12	17	2/12
	10	100	4/4	17	2/12	8	1/12	0	0/12	17	2/12
	11	100	4/4	17	2/12	33	4/12	42	5/12	17	2/12
	12	100	4/4	17	2/12	25	3/12	17	2/12	8	1/12
	13	100	4/4	0	0/12	17	2/12	0	0/12	0	0/12
HGD Il12b CGI 207-277	-895	100	4/4	0	0/12	0	0/12	0	0/12	0	0/12
	-866	100	4/4	0	0/12	17	2/12	8	1/12	0	0/12
	-846	100	4/4	0	0/12	8	1/12	0	0/12	0	0/12
	-836	100	4/4	0	0/12	0	0/12	0	0/12	0	0/12
	-833	100	4/4	0	0/12	8	1/12	0	0/12	0	0/12
	-804	75	3/4	33	4/12	50	6/12	33	4/12	25	3/12

		ESC clones with Pre-Methylated 601 + Il12b CGI 207-277 BAC									
		Pre-Meth BAC		Clone 1		Clone 2		Clone 3		Clone 4	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
601 Reduced CpG Insert	1	88	7/8	29	4/14	0	0/14	7	1/14	17	2/12
	2	86	6/7	50	7/14	0	0/14	14	2/14	33	4/12
	3	67	4/6	0	0/14	7	1/14	7	1/14	25	3/12
HGD Il12b CGI 207-277	-895	83	5/6	50	7/14	0	0/14	0	0/14	25	3/12
	-866	83	5/6	21	3/14	0	0/14	0	0/14	17	2/12
	-846	83	5/6	14	2/14	0	0/14	0	0/14	8	1/12
	-836	83	5/6	21	3/14	0	0/14	0	0/14	17	2/12
	-833	83	5/6	14	2/14	0	0/14	0	0/14	8	1/12
	-804	83	5/6	36	5/14	29	4/14	14	2/14	42	5/12

Figure 3-9 - Bisulfite Sequencing for CTCF Deletions and Tandem Arrays

A

ESC clones with Pre-Methylated Il12b CGI 15bp Deletion BAC

Il12b CGI Insert	Pre-Meth BAC		Clone 1		Clone 2		Clone 3		Clone 4		Clone 5		Clone 6	
	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
-1054	100	8/8	46	6/13	64	9/14	33	4/12	0	0/10	0	0/10	40	4/10
-1041	100	8/8	54	7/13	43	6/14	42	5/12	10	1/10	0	0/10	20	2/10
-1037	88	7/8	46	6/13	57	8/14	58	7/12	30	3/10	20	2/10	60	6/10
-990	88	7/8	50	7/14	31	4/13	42	5/12	10	1/10	10	1/10	30	3/10
-957	75	6/8	50	7/14	54	7/13	50	6/12	0	0/10	10	1/10	50	5/10
-938	100	8/8	50	7/14	54	7/13	42	5/12	20	2/10	20	2/10	30	3/10
-929	71	5/7	64	9/14	67	8/12	58	7/12	30	3/10	10	1/10	50	5/10
-927	86	6/7	50	7/14	67	8/12	50	6/12	0	0/10	10	1/10	50	5/10
-915	57	4/7	64	9/14	75	9/12	50	6/12	30	3/10	20	2/10	50	5/10
-910	86	6/7	64	9/14	92	11/12	50	6/12	20	2/10	20	2/10	50	5/10
-908	57	4/7	57	8/14	92	11/12	50	6/12	20	2/10	10	1/10	30	3/10
-898	86	6/7	57	8/14	75	9/12	50	6/12	30	3/10	0	0/10	30	3/10
-895	86	6/7	57	8/14	83	10/12	58	7/12	30	3/10	0	0/10	80	8/10
-866	86	6/7	57	8/14	67	8/12	42	5/12	0	0/10	10	1/10	60	6/10
-846														
-836														
-833														
-804	71	5/7	57	8/14	75	9/12	83	10/12	30	3/10	0	0/10	60	6/10

ESC clones with Pre-Methylated Il12b CGI 1-206bp BAC

Il12b CGI Insert	Pre-Meth BAC		Clone 1		Clone 2		Clone 3		Clone 4		Clone 5		Clone 6		Clone 7		Clone 8	
	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio	%	Ratio
-1054	88	7/8	7	1/14	21	3/14	7	1/14	43	6/14	0	0/10	20	2/10	30	3/10	22	2/9
-1041	88	7/8	21	3/14	43	6/14	43	6/14	29	4/14	0	0/10	30	3/10	0	0/10	44	4/9
-1037	88	7/8	7	1/14	36	5/14	14	2/14	36	5/14	10	1/10	30	3/10	10	1/10	56	5/9
-990	88	7/8	7	1/14	29	4/14	21	3/14	7	1/14	30	3/10	10	1/10	30	3/10	11	1/9
-957	88	7/8	7	1/14	21	3/14	29	4/14	7	1/14	30	3/10	40	4/10	10	1/10	33	3/9
-938	88	7/8	7	1/14	43	6/14	43	6/14	14	2/14	30	3/10	30	3/10	10	1/10	33	3/9
-929	88	7/8	14	2/14	29	4/14	36	5/14	36	5/14	10	1/10	40	4/10	40	4/10	67	6/9
-927	100	8/8	21	3/14	29	4/14	36	5/14	0	0/14	40	4/10	40	4/10	60	6/10	33	3/9
-915	88	7/8	7	1/14	29	4/14	0	0/14	43	6/14	50	5/10	40	4/10	30	3/10	56	5/9
-910	100	8/8	29	4/14	14	2/14	14	2/14	36	5/14	80	8/10	60	6/10	30	3/10	33	3/9
-908	100	8/8	7	1/14	14	2/14	14	2/14	7	1/14	30	3/10	20	2/10	20	2/10	33	3/9
-898	100	8/8	14	2/14	29	4/14	43	6/14	36	5/14	40	4/10	30	3/10	20	2/10	44	4/9

ESC clones with Pre-Methylated Il12b CGI 1-140 x2 BAC

		Pre-Meth BAC		Clone 1		Clone 2		Clone 3	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio
Il12b CGI 1-140	-1054	100	8/8	0	0/11	33	4/12	0	0/12
	-1041	88	7/8	45	5/11	25	3/12	58	7/12
	-1037	88	7/8	27	3/11	42	5/12	0	0/12
	-990	63	5/8	27	3/11	17	2/12	0	0/12
	-957	88	7/8	0	0/11	8	1/12	0	0/12
	-938	100	8/8	0	0/11	25	3/12	0	0/12
	-929	100	8/8	45	5/11	33	4/12	0	0/11
Il12b CGI 1-140 (2)	-1054	100	7/7	10	1/10	27	3/11	0	0/11
	-1041	100	7/7	0	0/9	27	3/11	0	0/11
	-1037	86	6/7	11	1/9	27	3/11	0	0/11
	-990	71	5/7	38	3/8	27	3/11	36	4/11
	-957	71	5/7	13	1/8	27	3/11	0	0/10
	-938	86	6/7	25	2/8	18	2/11	0	0/10
	-929	57	4/7	50	4/8	45	5/11	0	0/10

% of Methylated CpG
0
20
40
60
80+

ESC clones with Pre-Methylated 601x2 BAC

		Pre-Meth BAC		Clone 1		Clone 2		Clone 3	
		%	Ratio	%	Ratio	%	Ratio	%	Ratio
601	1	100	8/8	9	1/11	25	3/12	42	5/12
	2	100	8/8	27	3/11	17	2/12	25	3/12
	3	88	7/8	27	3/11	25	3/12	25	3/12
	4	88	7/8	27	3/11	8	1/12	17	2/12
	5	88	7/8	27	3/11	42	5/12	50	6/12
	6	88	8/8	45	5/11	42	5/12	42	5/12
	7	88	8/8	55	6/11	33	4/12	42	5/12
	8	88	7/8	18	2/11	17	2/12	17	2/12
	9	88	7/8	18	2/11	25	3/12	8	1/12
	10	75	6/8	45	5/11	33	4/12	25	3/12
	11	75	6/8	45	5/11	25	3/12	25	3/12
	12	88	7/8	27	3/11	42	5/12	42	5/12
	13	88	7/8	27	3/11	25	3/12	8	1/12
601 (2)	1	75	6/8	45	5/11	17	2/12	36	4/11
	2	100	8/8	27	3/11	8	1/12	45	5/11
	3	100	8/8	27	3/11	8	1/12	18	2/11
	4	100	8/8	18	2/11	8	1/12	9	1/11
	5	100	8/8	27	3/11	17	2/12	27	3/11
	6	100	8/8	82	9/11	42	5/12	36	4/11
	7	100	8/8	55	6/11	25	3/12	36	4/11
	8	100	8/8	27	3/11	17	2/12	18	2/11
	9	100	8/8	55	6/11	17	2/12	18	2/11
	10	88	7/8	27	3/11	33	4/12	9	1/11
	11	75	6/8	82	9/11	42	5/12	55	6/11
	12	88	7/8	55	6/11	67	8/12	55	6/11
	13	100	8/8	27	3/11	17	2/12	9	1/11

References

Bird, A. (1985). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG Island Mapping by Epigenome Prediction. PLoS Comput. Biol. *3*, e110.

Cole, H. a, Nagarajavel, V., and Clark, D.J. (2012). Perfect and imperfect nucleosome positioning in yeast. Biochim. Biophys. Acta *1819*, 639–643.

Davuluri, R. V, Grosse, I., and Zhang, M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. *29*, 412–417.

Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. *9*, e1003994.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., et al. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. *22*, 2399–2408.

Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. *4*, e1000138.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Gong, S., and Yang, X.W. (2005). Modification of bacterial artificial chromosomes (BACs) and preparation of intact BAC DNA for generation of transgenic mice. Curr. Protoc. Neurosci. *Chapter 5*, Unit 5.21.

Heintz, N. (2001). BAC to the future: the use of bac transgenic mice for neuroscience research. Nat. Rev. Neurosci. *2*, 1–10.

Kaplan, N., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2010). Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. Genome Biol. *11*, 140.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, a. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097.

Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. *276*, 19–42.

Macleod, D., Charlton, J., Mullins, J., and Bird, a P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282–2292.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Ramirez-Carrozzi, V.R., Nazarian, A. a, Li, C.C., Gore, S.L., Sridharan, R., Imbalzano, A.N., and Smale, S.T. (2006). Selective and antagonistic functions of SWI/SNF and Mi-2beta nucleosome remodeling complexes during an inflammatory response. Genes Dev. *20*, 282–296.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Vasudevan, D., Chua, E.Y.D., and Davey, C. a (2010). Crystal structures of nucleosome core particles containing the "601" strong positioning sequence. J. Mol. Biol. *403*, 1–10.

Wang, Z. (2001). An Efficient Method for High-Fidelity BAC/PAC Retrofitting with a Selectable Marker for Mammalian Cell Transfection. Genome Res. *11*, 137–142.

Warming, S., Costantino, N., Court, D.L., Jenkins, N. a, and Copeland, N.G. (2005). Simple and highly efficient BAC recombineering using galK selection. Nucleic Acids Res. *33*, e36.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res. *37*, W247–52.

Chapter 4

Determinants of the Unique Low DNA Methylation, Histone Modifications, and Nucleosome Occupancy at CpG Islands

Abstract

CpG islands are crucial sites of regulation in mammalian genomes, but it is currently unknown how they acquire their euchromatin features. To carefully analyze the role of dinucleotide content on chromatin changes, we created a dataset with all regions in the human genome with any CpG density above background. The CpG number, CpG density, and size of CpG rich regions correlated with CpG island features, but could not explain the extreme bias seen at promoter CpG islands for low DNA methylation, high H3K4me3, and DNase hypersensitivity. We found that the feature that correlates with both promoters and CpG island chromatin is transcription factor binding. Promoter CpG islands have extremely high transcription factor binding, which gives them their unique chromatin status. Only the feature of nucleosome density is solely affected by nucleotide content. Finally, we find evidence that CpG island regulation may be different in human and mouse cells.

Introduction

Epigenetic features can often be major determinants of gene expression and cell fate. One facet of epigenetics is the modification of cytosine-guanine dinucleotides (CpGs) by DNA methyltransferases, which covalently attaches a methyl group to C^5 of the cytosine base. This epigenetic mark is strongly correlated with repression and chromatin compaction (Klose and Bird, 2006). It is thought that the increased mutation rate of 5-methylcytosine has indirectly resulted in CpG depletion genome wide in mammals, as CpGs are four to five times less frequent than expected by random distribution (Shen et al., 1992).

Some sites exist where CpGs are protected from depletion. A unique feature of mammalian genomes is the accumulation of CpG dinucleotides at unusually high density, often referred to as a “CpG islands” (Bird, 1985; Gardiner-Garden and Frommer, 1987). CpG islands

are important sites of regulation in all cell types (Fouse et al., 2008). Strikingly, CpG islands can be found at 70% of coding gene promoters (Davuluri et al., 2001).

DNA methylation in mammals is the rule rather than the exception, with 60-80% average methylation genome wide in most cell types (Lister et al., 2009), but the CpGs within CpG islands often remain resistant to this modification (Suzuki and Bird, 2008). In addition, CpG islands are also strongly correlated with local modification of histone 3 lysine 4 with tri-methylation (H3K4me3), a nucleosome modification often found at transcribing genes (Mikkelsen et al., 2007; Santos-Rosa et al., 2002; Thomson et al., 2010). Study of large unmethylated CpG islands have found that they are often associated with the transcription machinery, even at intergenic sites (Illingworth et al., 2010). At some CpG islands, this mark is accompanied by deposition of a repressive mark, histone 3 lysine 27 trimethylation (H3K27me3), leading to a poised chromatin state termed a bivalent domain (Bernstein et al., 2006). Nucleosome depletion is another feature seen at CpG islands near promoters of active genes (Fenouil et al., 2012) These CpG island properties mediate chromatin accessibility which is thought to influence the speed of induction and the basal activity of island associated genes (Bhatt et al., 2012; Ramirez-Carrozzi et al., 2009). Regulation of all of these chromatin properties is unsurprisingly a complex process that is only recently beginning to be understood.

Research on CpG island properties has often considered each feature in isolation. Studies on CpG island DNA methylation have shown that many regions of low CpG methylation in the genome are controlled by discreet transcription factor binding sites (Lienert et al., 2011; Stadler et al., 2011). Histone modification does not seem to be driven by specific sites, as most CpG rich DNA is capable of acquiring H3K4me3 and H3K27me3 (Mendenhall et al., 2010), a process likely mediated by unmethylated CpG binding properties of proteins like Cfp1 (Thomson et al.,

2010). Nucleosome density at CpG islands may be dictated by nucleotide content; in vitro studies have shown that CpG rich DNA destabilizes nucleosome formation (Fenouil et al., 2012; Ramirez-Carrozzi et al., 2009). Understanding the interrelationship of these features remains a major goal of CpG island research.

Although common guidelines for CpG island definition have existed since their discovery, there is still no consensus. Many definitions use a sequence based approach with criteria adjusted to find the most promoter associated CpG islands and minimize repeat elements (Gardiner-Garden and Frommer, 1987; Saxonov et al., 2006; Takai and Jones, 2002). With the increasing availability of DNA methylation and chromatin datasets, some searches have been adjusted to define CpG islands based on favorable chromatin criteria (Bock et al., 2007; Fan et al., 2008). An experimental method for derivation of CpG islands, based on immunoprecipitation of Cfp1 associated DNA, demonstrated the need for flexibility in CpG island definitions. This study found that nearly half of all unmethylated CpGs islands are small and intergenic (Illingworth et al., 2010).

Answering fundamental questions about CpG island formation will increase our understanding of both transcription control at promoter proximal CpG islands and the function of promoter distal islands. In order to determine the nucleotide requirements for establishment of CpG island features, we created a dataset that includes every CpG-rich region in the genome, and then determined the chromatin state at each region. We found that many properties correlated with increasing CpG dinucleotide content, but significant enrichment for accessable chromatin features at CpG rich promoters is not explained by nucleotide content. The only chromatin feature strongly correlated with nucleotide content and not promoter location is nucleosome density, which we found to be largely influenced by GC content. Promoter CpG islands are

enriched for transcription factor binding, and the degree of enrichment correlates with level of CpG island features. This explains why promoters have a special chromatin environment compared to most intergenic CpG islands. Finally we compared CpG island regulation between human and mouse, finding that the mouse genome is more permissive to establishment of CpG island chromatin. Together are data suggest that the major determinant of CpG island chromatin is transcription factor binding.

Results

Defining All CpG Rich Regions in the Human Genome

In order to examine the effect of nucleotide content on the local chromatin environment, we developed a non-biased method to call Cytosine-Guanine dinucleotide rich regions using thresholds for minimum size and density. Unlike the most common Gardiner-Garden based criteria, our method does not consider GC content and computes regions purely based on CG occurance over 150bp windows. Our requirements were that (CpG #)/(Size * Probability of Random CpG, 1/16) >0.55. Using these criteria, we found 173,307 regions in the repeat masked human genome which we termed CpG Rich Regions (CGRs). CpGs are significantly depleted in the mammalian DNA, but clusters of CpG rich DNA make up a larger than expected portion of the genome (Table 1). Our criteria for CGR calling was left permissive in order to better understand the properties that lead to CGI formation. Even though the majority of regions are of minimal CpG number (6 CpGs, 150bp), a large dynamic range exists for CpG number, size, and CG density (Fig 4-1). Notably, our CGRs overlap the UCSC CpG Island dataset almost entirely, the only exceptions being portions of UCSC CGIs overlapping repeat DNA.

The majority of CGRs have CpG density at or near the minimal criteria, but nearly a third of CGRs have high CpG number, density, and size attributes (Fig 4-1a-c). As expected, GC content is higher than the genomic average of 45-50% with a median range of 55-60% (Fig 4-1d). The total CpG number in a CGR is closely linked to the CGR size (Fig 4-1e). Interestingly, no regions in the genome are maximally CpG enriched. CpG rich DNA is thought to be structurally rigid and may be evolutionarily constrained to prevent breakage. As such, the highest density seen in the human genome is 2.4 Obs/Exp, or about 30% CpG content, and the vast majority of CGRs remain under 20% CpG content. Unexpectedly, total CpG counts at CGRs are often higher than theoretically necessary at CGRs with sizes over 1kb (Fig 4-1e). The bulk of large sized CGRs exceed the CpG number necessary for discovery, suggesting large CGRs may be under selective pressure to increase CpG density. Maximum GC% of CGRs increases with CpG density, although CpG content is not constrained by GC% (Fig 4-1f). Low CpG density regions may occur in high GC % patches and vice versa. For example, approximately 3300 CGRs with high CpG density (above 0.75) have GC content in the 45-55% range.

CpG Rich Regions Are More Common and More Pronounced at Promoters

Distribution of CGRs across chromosomes is unequal, but is correlated with chromosome size; chromosome 1 has the greatest number of CGRs and chromosome Y the least (Fig 4-2a). However, even after normalizing for chromosome size, chromosomes 15, 16, 19, and 21 are still aberrantly CGR enriched (Fig 4-2b). This unusual CGR frequency is explained by the higher coding gene density on these chromosomes (Fig 4-2c), which strongly suggests that CGRs are constrained to form near genes.

In affirmation of this precept, we find that nearly two thirds of all CGRs are associated with genes (Fig 4-3a). Half of gene associated CGRs lie within introns, but the remainder, over one third of all CGRs, lie directly on promoters, UTRs, or coding exons. This is especially remarkable considering that coding DNA makes up less than 2% of the genome. Our criteria for promoter proximal CGRs was that there must be overlap within 500bp of a transcription start site. By this criteria, CGRs are proximal to the promoters of 15,876 protein coding genes, or just over two thirds of all coding genes. CGRs are also near the start sites of over 4000 non-coding transcripts. Distribution of promoter CGRs to promoters is not always 1:1, as many genes have multiple CGRs at alternative transcription start site locations or share a CGR with another gene. Approximately 1250 pairs of coding genes share a single CGR, utilizing a bidirectional promoter.

To investigate whether CGRs have different properties based on gene proximity, we analyzed the CpG number, CpG density, size, and GC % for each CGR after classifying them by genomic location (Fig 4-3b, Fig 4-4). Strikingly, among all genomic locations only promoter CGRs were substantially enriched for CpG density. Over 80% of CGRs at promoters are significantly enriched for moderate to high CpG density. Promoter CGRs are also uniquely enriched for total CpG number and CGR size (Sup Fig. 2b,c). High GC content was not as remarkably enriched at promoters, but promoter CGRs were still 2-4 times more likely to have high GC content than other regions (Fig 4-3c). To determine whether CpG dense CGRs were constrained to promoters, we classified every CGR by density and calculated the fraction within each genomic location. Over 66% of high density CGRs are at promoters. The next most frequent CpG dense location is intergenic, making up a substantially lower 16% of high CpG density CGRs. Similarly, CGRs classified as high in CpG number or size are also most likely to

be located at promoters (Fig 4-4d,e). While high density CGRs are not exclusively promoter based, they are extremely likely to be located at promoter. Taken together, our analysis of CGRs by genic location finds that promoter CGRs are uniquely enriched for CpG dinucleotides.

CGRs Are Enriched For Euchromatin Modifications

Using publicly available datasets we were able to determine the chromatin state at CGRs in several different cell types, but with principal interest in human embryonic stem cells. Embryonic stem cells have not yet undergone lineage commitment and contain many chromatin features that suggest a less repressive state, such as the presence of bivalent domains and reduced heterochromatin foci (Meshorer et al., 2006). Therefore, they may provide the best system for studying establishment of chromatin features at CpG islands. We analysed DNA methylomes, histone modification ChIP-Sequencing datasets, and nucleosome occupancy datasets.

DNA methylation at CGRs is extremely bimodal. In all cell types assayed, the vast majority of CGRs are marked with either very high of very low CpG methylation (Fig 4-5a). In embryonic stem cells, most CGRs with low DNA methylation have under 5% of their CpGs methylated. We considered low DNA methylation to be below 30% averaged across a CGR. We found that 28,897 CGRs, or 17% of all CGRs, had low DNA methylation in ES cells. The average DNA methylation for CGRs with high methylation is generally more variable, extending over a range of 75%-100%. We find that approximately 30,000 CGRs have H3K4me3 modification at a level above background in ES cells, distributed in an approximately linear fashion (Fig 4-5b). We considered two metrics for CGR nucleosome occupancy; DNase hypersensitivity, and nucleosome density derived from MNase sequencing. DNase hypersensitivity peaks described in ES cells overlap with 36,200 CGRs, with intensity scores

linearly distributed (Fig 4-5c). In a MNase-seq dataset from K562 cells, average nucleosome density at CGRs has a bell-shaped curve distribution (Fig 4-5d). We were limited a non-ES cell line by availability, but the nucleosome occupancy data in K562 leukemia cells is highly similar to another ENCODE MNase-seq dataset from GM12878 cells, and should still reflect intrinsic nucleotide stability effects. Approximately 46,000 CGRs have low nucleosome density.

Next we investigated the interrelationship of these chromatin features at CGRs. The only obligate dependency between features that we observed is the requirement for CGRs to have high H3K4me3 signal in concert with low DNA methylation (Fig 4-5e). Nearly all (98%) of CGRs with high H3K4me3 signal also have low DNA methylation in ES cells. Notably, this obligate relationship is unidirectional; only 37% of all low methylated CGRs have a high H3K4me3 signal and 17% have no H3K4me3 signal. Still H3K4me3 and DNA methylation are the most closely linked modifications, with a correlation coefficient (PCC) of -0.72, where +/-1 is perfect correlation. DNase hypersensitivity sites are present at 88% of CGRs with high H3K4me3 and 70% of CGRs with low DNA methylation. Additionally, nucleosome density in K562 cells remains roughly evenly distributed for each set of CGRs classified by chromatin, indicating that no chromatin properties significantly enrich for nucleosome depletion. Interestingly, H3K27me3 is preferentially deposited at CGRs that also have low DNA methylation (PCC = -0.33) (Fig 4-6b). However, there is no intercorrelation between H3K27me3 and either H3K4me3 or DNase hypersensitivity.

In order to determine which proportion of CGRs could be considered full featured CpG islands, we calculated the frequency of chromatin feature overlap. 23909 of the 28897 CGRs with low DNA methylation also have moderate to high H3K4me3 signal (Fig 4-6a). 20,000 of these CGRs overlap a DNase hypersensitivity peak, but only 10,000 have low nucleosome

density in MNase-seq. Approximately a third of the unmethylated, H3K4me3 marked CGRs also have appreciable H3K27me3 signal and could be considered bivalent domains. Presense of H3K27me3 does not change the distribution of nucleosome occupancy at CpG island-like CGRs.

Promoter CGRs are Uniquely Enriched for CpG Island Chromatin Features

Previous research has largely focused on features at promoter CpG islands, so we wanted to see how promoter CGR chromatin modifications differed from other genomic locations. Analyzing ES cell CpG methylation across all CGRs reveals that promoter CGRs are over six times more likely to have low average DNA methylation than CGRs at other genomic locations (Fig 4-7a). This unmethylated promoter CGR bias is nearly identical in frontal cortex tissue and periphral blood leukocyte methylomes (Fig 4-8c,d). Across all CGRs, 53% of low DNA methylated CGRs are at promoters (Fig 4-7e). Besides promoters, 13% of intergenic CGRs have low DNA methylation, while only 3% of CGRs at exons have low DNA methylation. This equates to low DNA methylation at 15,332 promoter CGRs, 7373 intergenic CGRs, and 904 exonic CGRs (Fig 4-8b). CGRs at promoters are also enriched for H3K4me3 signal and DNase hypersensitivity peaks (Fig 4-7b,c). 76% of promoter CGRs are marked by medium to high H3K4me3. H3K4me3 signal is extremely biased towards promoter CGRs, as 78% of high H3K4me3 marked CGRs are located there. DNase hypersensitivity signal is also strongly enriched at promoters, with over 73% of promoter CGRs overlapping some DNase-seq signal and inversely over 53% of DNase-overlapping CGRs located at promoters. The majority of promoter CGRs did not contain high H3K27me3, but this repressive mark is still more common at promoter CGRs than at other genomic locations (Fig 4-8e).

Interestingly, CGR promoter location also correlated with nucleosome density, a feature which previously did not correlate with any other chromatin property. At promoter CGRs there is a 1.5-2x enrichment for low nucleosome density, and a similar fold depletion for high nucleosome density, compared to CGRs at other genomic locations (Fig 4-1d). The lack of correlation between nucleosome density and the other chromatin features enriched at promoters suggests that these features may be controlled by different mechanisms. Taken together, most promoter CGRs clearly have a striking enrichment for CpG island features compared to CGRs at other genomic locations.

Comparing promoter and non-promoter CGRs with full CpG island features, we find that two thirds of full featured CGRs are located at promoters (Fig 4-8a). Bivalent CGRs are even more likely to occur at promoters than non-promoter locations. 6515 promoter CGRs have low DNA methylation and at least moderate H3K4me3 and H3K27me3, compared to only 1772 non-promoter CGRs.

Promoter Location More Important than Nucleotide Content for CpG Island Features

Next, we wanted to determine to what extent the unique chromatin environment at promoter CGRs could be explained by the underlying nucleotide composition. Since promoter CGRs are enriched for CpG density compared to non-promoter CGRs, we normalized for CpG density by dividing promoter and non-promoter CGRs further into two classes; CGRs above (high) and below (low) our cutoff for medium CpG density (0.75). CGRs in the high CpG density class have similar nucleotide properties between promoter and non-promoter regions (Fig 4-10a). The high CpG density class also has similar amounts of CGRs between locations, with 15,201 high density promoters regions and 18,639 high density non-promoter regions. Analyzing

the DNA methylation at CGRs in the high CpG density class reveals that promoter CGRs are nearly three times as likely to have low average DNA methylation compared to non-promoters (Fig 4-9a). 90% of high density promoter CGRs have low DNA methylation in ES cells compared to 33% of high density non-promoter CGRs. A stark difference also exists between low CpG density CGRs separated by genomic location; 33% of the low density promoter CGRs possess low DNA methylation, which is over five fold more frequent than low density non-promoter CGRs. Comparing the high CpG density to low CpG density classes shows that nucleotide properties do influence DNA methylation, but does not explain the disparity between promoters and non-promoters.

The H3K4me3 modification is also unusually enriched at promoters after normalizing for CpG density; 91% of high density promoter CGRs have H3K4me3 signal in contrast to only 32% of non-promoters (Fig 4-9b). DNase hypersensitivity is also enriched at high CpG density promoters, with 89% of high density promoter CGRs overlapping DNase sites in contrast to 41% of high density non-promoters CGRs (Fig 4-9c). Promoter CGR feature bias is still present at even more stringent nucleotide cutoffs. When promoter and non-promoter CGRs are separated by a combined CpG density and a CG number cutoff, we still see a large difference between promoters and non-promoters for DNA methylation (Fig 4-10b). For the CpG island features of low DNA methylation, H3K4me3 and DNase hypersensitivity, there is a strong prejudice towards enrichment at promoter CGRs that is not explained by underlying CpG content.

Interestingly, nucleosome density is normalized between promoter and non-promoter CGRs by CpG density (Fig 4-9d). The same proportion of high CpG density promoter CGRs and high density non-promoter CGRs had low nucleosome occupancy, approximately 47%. This correlation between nucleosome density and CpG density is evidence that nucleotide content is a

determinant of histone positioning. To define the nucleotide feature that had the largest effect on local histone density, we determined correlation coefficients for nucleosome density paired with basic CGR properties such as CpG number, density, and GC content (Fig 4-9e). Surprisingly, GC content had the strongest correlation with low nucleosome density. CpG number and density are the next strongest correlates, speculatively because of their intrinsic effect on GC content. To demonstrate the effect of GC content on nucleosome density at CGRs we divided all CGRs into high (60%+) and low (<60%) GC content classes. We chose 60% because it is roughly the mean GC content for all CGRs. When looking at the nucleosome density for high and low GC content classes we see a striking enrichment for nucleosome depletion in the high GC class (Fig 4-9f). Approximately two thirds of all CGRs with low nucleosome density have a GC content greater than 60%. A slightly stricter threshold of 65% GC content strongly enriches for low nucleosome density. CGRs with 65% or greater GC content are 4 times more likes than low GC % CGRs to have low nucleosome density, and are 4 times more depleted for high nucleosome density. GC content clearly has a destabilizing effect on nucleosome density in the human genome at CpG rich regions.

Normalizing for high CpG density also removes the promoter bias for H3K27me3 (Fig 4-10d). In contrast to H3K4me3, approximately $1/5^{th}$ of both high CpG density promoter CGRs and non-promoter CGRs have a high H3K27me3 signal.

Transcription Factor Binding is a Unique Property of Promoter CGRs which Contributes to CpG Island Feature Establishment

In order to understand what the source of the unique chromatin at promoter CGRs could be, we decided to investigate transcription factor binding to DNA. We considered 132 CGR-

binding transcription factors(TFs) in the ENCODE ChIP-seq database, allowing data from every available cell type. Approximately a third of CGRs accounted for all TF binding (Fig 4-12a). For each TF bound CGR, we summed all TF peak scores and classified the aggregate score as weak binding, substantial binding, or high (hotspot) binding. Promoters CGRs are strongly enriched over all other genomic locations for high TF binding (Fig 4-11a). Similar to what was previously seen with several chromatin features, normalizing for CpG density between promoters and non-promoters does not explain TF binding enrichment (Fig 4-11b). After applying a high CpG density cutoff, promoter CGRs are still 4 times more likely than non-promoter CGRs to have high TF binding, and the majority of promoter CGRs (70%) are hotspots.

To investigate whether the level of TF binding could account for low DNA methylation at promoters, we looked at the average DNA methylation in ES cells for CGRs classified by TF binding score (Fig 4-11c). Hotspot TF binding CGRs are strongly enriched for low DNA methylation, and there is a linear relationship between the TF binding score and DNA methylation. TF binding's relationship to low DNA methylation is not ES cell specific; 87% of hotspot binding CGRs have low DNA methylation in periphral blood leukocytes and frontal cortex (Fig 4-12b, data not shown). Hotspot TF binding is also strongly correlated with H3K4me3 signal at all CGRs (Fig 4-11c). Moderate to high H3K4me3 can be found at 74% of hotspot binding CGRs (Fig 4-11d). Like low DNA methylation, the proportion of CGRs with H3K4me3 drops off sharply with decreases in TF binding. The TF aggregate score alone does not completely explain the chromatin properties at CGRs, as 12% of hotspot CGRs retain high DNA methylation and 26% have no appreciable H3K4me3. To some extent this is expected, as transcription factor binding may have a myriad of different functions. However, the intensity of binding at promoter CGRs strongly correlates with the presense of CpG island features.

In order to understand more about the hotspot landscape, we looked at the identity of individual transcription factors bound to CGRs. The top CGR binding protein is RNA Polymerase II (Fig 4-11e), unsurprising as the transcriptional machinery defines promoters and was shown to associate with most large unmethylated CpG islands (Illingworth et al., 2010). The highest CGR binding proteins also includes general transcription machinery and ubiquitous transcription factors. Observing the chromatin at CGRs specifically bound by individual transcription factors suggests possible roles for some factors in the establishment of CpG island features (Fig 4-11f). In general, nearly every TF analyzed has low average DNA methylation, a high DNase hypersensitivity peak and a high TF aggregate score across all CGRs they bind. CTCF and the Rad21 cohesin are closesly associated genomic organization proteins that bind CGRs with similar properties. They both prefer less dense CGRs and also are bound to regions with weaker H3K4me3 signal, suggesting a promoter distal role that mostly affects DNA methylation. The transcription machinery and general transcription factors are commonly bound at CGRs with high CpG density as well as high GC content, complete with corresponding nucleosome depletion. The only ENCODE transcription factors strongly associated with H3K27me3 were Suz12, a component of the Polycomb Repressor Complex 2, and CtBP2, a common repressor. Both are frequently bound to CGRs, prefer high CpG density, and their presense antagonizes H3K4me3. Suz12 is unique among all ENCODE factors analyzed as the CGRs bound by Suz12 have by far the lowest TF aggregate score, suggesting binding leads to exclusion of other transcription factors. Tissue specific transcription factors also bind many CGRs; the ES pluripotency network protein Nanog binds low DNA methylated CGRs with high H3K4me3 in ES cells, while Gata2, a protein important in hematopoietic cells, binds CGRs which have moderate DNA methylation in ES cells and a relatively diminished H3K4me3 signal.

Some tissue specific factors like Nanog, NFκB and Egr-1 seem to prefer large, CpG dense CGRs, while others like Pou5f1(Oct4) and Gata-2 are more commonly bound to less dense CGRs and may be more important at promoter distal enhancer regions. A few ENCODE TFs prefer to bind CGRs with high nucleosome density; GCN5, HDAC8, and BRF2. These proteins collectively bind only 333 CGRs and therefore are not a major factor in CpG island regulation.

The collective binding properties of the ENCODE transcription factors remain consistent when only considering ChIP-Seq data from H1 ES cells (Fig 4-12c). With few exceptions almost every TF binds CGRs with high CpG density, low DNA methylation, low nucleosome occupancy, high H3K4me3, and a high TF aggregate binding score. TF binding is strongly correlated with CpG island features, and analysis of individual TFs finds CGR properties are related to known functional roles. We therefore conclude that transcription factor binding is the most important determinant of the uniquely low DNA methylation and high H3K4me3 signal at promoter CpG islands. Interesting, most TF factors tend to prefer GC rich low nucleosome occupancy CGRs as well, suggesting that GC content may be a sequence determinant that promotes TF binding.

A TF driven model is compelling, as it explains the lack of a minimum threshold for acquisition of chromatin features. As an example, the most common CGRs are 6 CpG regions approximately 150bp in size, the majority of which have high DNA methylation. However, 761 of these 6 CG regions maintain low DNA methylation in four different cell types (Fig 4-13a). Many of these CGRs share the chromatin properties of another CGR within 500bp, but the remaining 277 are isolated. The low methylation at these 277 CGRs can be almost completely explained by binding of a single ENCODE transcription factors, CTCF, which is found at 223 of

these CGRs. In addition, protection from DNA methylation by direct TF binding could explain the often heterogeneous methylation state at the edges of CpG islands (Fig 4-13b).

Differences Between Human and Mouse CpG Island Regulation

Interested in CGR conservation between the two most common model systems, we defined CpG Rich Region mouse genome using the same method. The mouse genome contains an approximately equal amount of CGRs with similar average CpG density to the human dataset, but mouse CGRs are also about 25-30% smaller on average (Fig 4-14a). Additionally, CGRs in mouse ES cells have approximately 10% lower average DNA methylation compared to human ES CGRs. The distribution of CpG density within all mouse CGRs seems to reflect the size disparity to human CGRs, as 19% of human CGRs are moderately dense or higher compared to 12% for mice (Fig 4-14b). Distribution of DNA methylation at all CGRs is also highly similar (Fig 4-14c), with 28,897 and 34,107 CGRs with low DNA methylation in humans and mouse respectively. However, normalizing for CpG density reveals a stark difference between the two species: 37% of high CpG density CGRs in human ES cells have high DNA methylation, but only 5% of high CpG density mouse CGRs have high DNA methylation (Fig 4-14d).

In order to decipher whether the protection of CpG dense CGRs from DNA methylation in murine ES cells was due to species differences or cell type differences, we examined the average DNA methylation at murine CGRs in peritoneal macrophages and adult frontal cortex tissue. As in murine ES cells, the proportion of high CpG density CGRs with high DNA methylation remains much lower than the comparable class in human ES cells (Fig 4-14e). Murine CGRs are consistently more likely to have low DNA methylation at regions of high CpG density.

To compare promoter CGRs across species, we generated a subset of human and mouse orthologues containing 12,674 promoter CGRs found at the same gene in both species. The distribution of DNA methylation at the orthologous CGRs was similar across species in both ES cells and in frontal cortex tissue (Fig 4-15). Presumably, human and murine ES cells will methylate similar CpG island promoter targets to silence genes antagonistic to pluripotency. However, when the 773 human promoter CGRs with high ES cell DNA methylation are compared to their mouse orthologues, we find that high DNA methylation is only conserved in 25% of murine CGRs (Fig 4-14f). In mouse ES cells, high CpG density at a promoter CGR overides cross species conservation of a high DNA methylation state 99% of the time. This effect transcends cell type, as we found that in human and mouse frontal cortex samples only 50% of mouse orthologuous to methylated human CGRs are also methylated. As seen in ES cells, only 20% of the high CpG density murine orthologues recapitulate the high methylation at human CGRs in frontal cortex cells. In contrast to human cells, murine CGRs may use nucleotide content as a major determinant of DNA methylation state.

Discussion

By including any region in the genome with even slight CpG density in our analysis of human CpG islands, we have completed an unprecendented description of the effect of nucleotide content on CpG island properties, and of the relationship between CpG island chromatin features. This data allowed us to discern basic interrelations such as the frequent presense of DNase at H3K4me3 marked CGRs and the preference for H3K27me3 deposition at moderately H3K4me3 marked CGRs. CpG island chromatin features increase in strength along with CpG content, CpG density, and size. However, this one step model did not explain all CpG

island chromatin. For instance, H3K4me3 clearly requires unmethylated CpGs , but unmethylated CpGs are not sufficient for H3K4me3 recruitment. Similarly, we found no consistent thresholds or cutoffs for stable establishment of low DNA methylation at CGRs. We could only find a basic correlation to CGRs with high CpG content.

However, CGRs with high CpG content, density and size are also nearly always found at gene promoters. By comparing promoter and non-promoter CGRs with similar nucleotide properties, we found that genomic location was a more important determinant of CpG island feature establishment than the underlying nucleotide content. This was true for the CpG island features most closely associated with active euchromatic regions; low DNA methylation, high H3K4me3, and high DNase hypersensitivity.

To determine what made promoter CpG islands unique, we examined transcription factor binding and found that promoter regions have extreme TF binding activity compared to non-promoter regions. This is not suprising in and of itself, but TF binding is also strongly correlated with CpG island chromatin features. We therefore propose that the major determinant of the acquisition of euchromatin features is the binding of TF factors, for both promoter and intergenic CpG islands.

A model based on TF driven CpG island features does not rule out contribution from nucleotide content. Indeed, the antagonistic relationship between GC content and nucleosome occupancy may contribute to chromatin accessibility at some CpG islands. Any factors which increase accessibility will likely increase functional transcription factor binding, a paradigm seen at the promoters of inducible genes (Ramirez-Carrozzi et al., 2009).

Another clear role for CpG content is the importance it has in protein-DNA recognition. The top CGR binding proteins included Sp1 and CTCF, which both have CpGs in their

consensus binding motifs and are both implicated in demethylation of binding targets (Holler et al., 1988; Kim et al., 2007; Macleod et al., 1994; Stadler et al., 2011). If CpGs are present in certain binding motifs, increasing CpG content at CpG islands may increase TF binding. Note that in this case, the nucleotide content is secondary to the nucleotide order, which could explain why nucleotide content alone is not sufficient to predict chromatin properties. Evidence for enrichment of particular binding sites in CpG islands comes from a computational examination of CpG island sequences which extracted species specific patterns (Chae et al., 2013). These patterns can be used to rebuild the examined species phylogeny, suggesting they may be the result of evolving binding sites. This model could also explain the conservation of large CpG islands at housekeeping genes, as the immense size of certain CpG islands may stochastically increase TF binding sites and therefore stabilize euchromatin formation.

Finally, we also demonstrated that CpG island regulation differs between species. As opposed to human cells, in murine cells DNA methylation has a pronounced sensitivity to CpG density, as nearly all CpG dense CGRs have low DNA methylation. As the CpG islands are smaller in the mouse genome, they may require a different demethylation mechanism to maintain their regulative activity. Another possibility is that the transcription factor based demethylation mechanism has evolved abnormal activity in mice, leading to a relaxation of evolutionary constraints on CpG island size. Whatever is responsible for this CpG threshold effect, we also see that it seems to be stronger in a pluripotent cell type. ES cells are known to have unique chromatin environments (Meshorer et al., 2006; Xu et al., 2009), but it remains to be seen how this will inform our understanding of the differences between human and mouse CpG island chromatin.

Materials and Methods

Derivation of CpG Rich Regions

A script was created to scan the hg19 human genome, downloaded from UCSC Genome Browser (Kent et al., 2002), for CpG occurance. For sliding 150bp windows, regions that qualified for CpG Density > 0.55 were kept. CpG Density = (CpG #)/(Size * Probability of Random CpG, 1/16). Overlapping regions were integrated. The output was genomic coordinates of all CG rich regions across all 23 chromosomes.

For mouse CGRs, we used the same script and conditions on the mm9 mouse genome, downloaded from UCSC Genome Browser.

Analysis of CGR CpG Properties and Genomic Location

Analysis was primarily done using frequency histograms with predetermined bins and with correlate functions in Microsoft Excel. Genomic location of non-promoter regions was extracted from the UCSC Refseq table database. Promoter location was determined by overlap +/- 500bp of any hg19 transcription start site. Overlapping locations were called by this heiarchy: promoter, UTR, exon, intron, intergenic.

Analysis of CGR Chromatin Properties

H1 ESC H3K4me3 and H3K27me3 ChIP-sequencing datasets were obtained from the publically available ENCODE database on UCSC Genome Browser (Bernstein et al., 2012; Ernst et al., 2011). The histone modification signal at CGRs was calculated by averaging of the ChIP sequencing score across the CGR interval. H1 ESC DNase Hypersensitivity sequencing and MNase Sequencing data in K562 cells were also downloaded from the ENCODE database on

UCSC Genome Browser. DNase signal at CGRs was calculated as the maximal peak score overlapped by the CGR. MNase nucleosome density was calculated by averaging the MNase sequencing signal over the CGR.

DNA methylation at CGRs was processed from published datasets. The ES cell bisulfite sequencing methylome data is from HUES64 cells, and the frontal cortex data is from patient biopses (Ziller et al., 2013). Periphral blood lymphocytes were obtained by Ficolli gradient centrifucation and hair follicles were picked from male patients as described for bisulfite sequencing methylomes (Kunde-Ramamoorthy et al., 2014). Mouse methylomes were obtained from bisulfite sequencing in ES ROSA V6.5 cells (Vincent et al., 2013) and adult frontal cortex samples (Lister et al., 2013).

Transcription factor ChIP-sequencing data was taken from the human ENCODE Uniform Peaks database. Calculation of all cell-type binding was done by overlap of peaks in version 2 of the TF database with all CGRs. Calculation of H1 ESC TF binding only was done by sorting of version 3 for H1 ESC entries and overlap of peaks with all CGRs. Each overlap was calculated in two ways: 1- For each individual TF, all CGRs it binds. 2- For each CGR, all TFs that bind. For this calculation, all binding TF ChIP-Seq peak scores were summed to yield the TF aggregate binding score at each CGR.

Figure Legends

Table 1 : General Statistics for Human CGRs – Coverage, averages, and maximums for all human CGRs.

Figure 4-1- Nucleotide Properties of CpG Rich Regions – Frequency of the 173307 CGRs within defined ranges of CpG number(A), CpG Density (B), Size (C) or GC content (D). Intercorrelations between these properties are displayed as frequency histograms of Size bins and CpG number bins (E) or CpG Density bins and GC content bins (F). Red indicates high occurance, white indicates low occurance, and grey represents theoretical limits.

Figure 4-2 - CGR Location by Chromosome – Human CGR distribution across chromosomes (A), or normalized by chromosome size (B) or number of coding genes (C). Only protein coding genes are represented.

Figure 4-3 – Promoters are Enriched for CpG Dense Regions – (A) Frequency of CGRs at each genomic location. Promoters are +/- 500bp from an annotated TSS. (B) Fraction of each genic class that has high medium or low CpG density. Enrichment is displayed as % of total CGRs in each respective genic class. (C) All CGRs were classified as high, medium, or low CpG density; the fraction of each identified by genic location is shown. (D) CGRs are separated into CpG Density classes and then fraction of class in each genomic location is shown

Figure 4-4 - **Promoters are Enriched for High CG Number and Large CGR Size** (A) Raw histogram of all CGRs, separated by genic location and CpG density. Red colors high frequency,

white marks low occurance. (B) Fraction of CGRs in each CpG number class, high medium or low, for each genomic location. (C) Fraction of CGRs in each Size class, large medium or small, for each genomic location. (D) Fraction of CGRs in each genomic location, for each CpG number range. (E) Fraction of CGRs in each genomic location, for each Size range. Genic location was determined by overlap within 500bp of an annotated transcription start site for promoters, or any overlap with another Refseq defined feature.

Figure 4-5 – Description of Chromatin Environment and Correlations at CpG Rich Regions – Graphs depict CGR frequency across the score or signal for each chromatin property-average DNA methylation in ESCs(A), average ESC H3K4me3 signal (B), high ESC DNase hypersensitivity score (C), and nucleosome density at CGRs in K562 cells (D).
(E) Intercorrelation between chromatin features. CGRs are subdivided by chromatin class at top then chromatin property at left is displayed as percent distribution of CGRs within the class. The correlation coefficient for each pair of features is displayed below the distribution graph.

Figure 4-6 – CGRs with CpG Island Properties

(A) Frequency of CGRs which share certain chromatin features. LowM=Low DNA Methylation, K4m3= High H3K4me3, DNase=Overlap with a DNase hypersensitivity peak, LNucD= Low Nucleosome Occupancy, K27m3= High H3K27me3. Note that DNase Hypersensitivity and Low Nucleosome Occupancy are neither mutally exclusive nor inclussive (see Fig 4-5e). All datasets are from human ESCs except for Nucleosome Occupancy, which is from K562 cells

(B) Intercorrelation between H3K27me3 in ES cells and chromatin features. Row 1 is chromatin distribution for all high H3K27me3 CGRs, row 2 is H3K27me3 distribution within the other chromatin classes

Figure 4-7 – Promoters CGRs are enriched for CpG Island Features Percent of total CGRs in each genic class separated by (A) average DNA methylation in ESCs across each CGR (B) Low, medium or high H3K4me3 ChIP-Seq signal in ESCs across each CGR(C) Overlap of each CGR with DNase hypersensitivity peaks from ESCs or (D) average nucleosome occupany in K562 cells across each CGR. (E) Percent of total CGRs classified by high CGI chromatin property classes that lie within each genic location, as in Fig 4-3d.

Figure 4-8 – DNA Methylation at Promoters, in Different Cell Types, and Other Chromatin Modifications (A) Count of CGRs which share certain chromatin features, separated by genomic location. LowM=Low DNA Methylation, K4m3= High H3K4me3, DNase=Overlap with a DNase hypersensitivity peak, LNucD= Low Nucleosome Occupancy, K27m3= High H3K27me3, as in Fig 4-6b (B) Frequency histogram of total CGRs in each genic class separated by averaged DNA methylation in ESCs (C) Percent of total CGRs in each genic class separated by average DNA methylation in a frontal cortex tissue methylome across each CGR (D) or by average DNA methylation in a periphral blood leukocytes methylome. (E) Percent of total CGRs in each genic class with H3K27me3 ChIP-Seq signal across each CGR

Figure 4-9 – Effect of Nucleotide Composition and Genomic Location on Chromatin at CpG Rich Regions

Comparing (A) average ESC DNA methylation at promoter and non-promoter CGRS, separated by CpG density (above and below 0.75). (B) average H3K4me3 signal between pro/non-promoter CGRs with high/low CpG density (C) DNase hypersensitivity peak overlap between pro/non-promoter CGRs with high/low CpG density (D) nucleosome occupancy in K562 cells between pro/non-promoter CGRs with high/low CpG density (E) Graph of the correlation between nucleosome occupancy and nucleotide or chromatin features (F) Nucleosome occupancy at CGRs subdivided into low GC%, mid to high GC%, or high GC%.

Figure 4-10 – Effect of Higher CpG Count and Density Cutoffs on Chromatin at CpG Rich Regions (A) Comparing total CpG number at promoter and non-promoter CGRs, separated by CpG density (above and below 0.75). (B) Frequency histogram comparing average ESC DNA methylation at promoter and non-promoter CGRs, separated by high/low CpG Density (C) Comparing average ESC DNA methylation at promoter and non-promoter CGRS as in (B), but with stricter nucleotide criteria – High CpG density and high CpG number indicates over 0.75 CpG Density and over 50 total CpGs in a CGR. (D) average ESC H3K27me3 at promoter and non-promoter CGRS, separated by CpG density

Figure 4-11 – Transcription Factor Binding Enrichment at Promoters Affects Chromatin Features – (A) ENCODE TF binding aggregate score classes, separated by genic location. Shown is % of genic location with no binding, weak binding, substantial binding, or high TF binding hotspot (B) Percent of promoters and non-promoter CGRs in each TF aggregate score

class, subdivided by high and low CpG density CGRs (0.75) (C) Percent of TF Aggregate Score class with low medium or high DNA methylation across CGRs (D) Percent of TF Aggregate Score class with low medium or high H3K4me3 ChIP-Seq signal across CGRs (E) Top 20 ENCODE TFs bound to the most CGRs in any cell type (F) CpG island feature heatmap for a selection of functionally grouped Transcription Factors, with average nucleotide and chromatin properties for the CGRs they bind. Red on the heatmap indicates a higher absolute value, blue indicates a lower absolute value. ES= Embryonic Stem Cell HF= Hair Follicle PBL=Periphral Blood Leukocyte Nuc=Nucleosome

Figure 4-12 – Transcription Factor Binding at CGRs in Human ES cells

(A) Aggregate transcription factor binding at all CGRs, across all cell types available in the ENCODE TF database, as a percent of all 173307 CGRs (B) Percent of CGRs classified by TF binding with low, medium, or high DNA methylation in the periphral blood leukocyte methylome. (C)For all TFs from ENCODE with human ES cell ChIP-Seq data, shown are heatmaps for the average properties of overlapped CGRs. Arranged from TFs with most CGRs bound to TFs with least

Figure 4-13 – Evidence of TF Binding Effects on CGR Chromatin

(A) Count of the minimum 6CpG CGRs which are unmethylated in 4 different cell types, 6CpG CGRs that are not within 500bp of another CGR, and count of those CGRs that also have CTCF binding. (B) Visualization in UCSC Genome Browser of a large CGR at the LPAR5 exon, with Frontal Cortex DNA methylation and ES methylation displayed below. Beneath the methylation tracks are ENCODE TF Binding peaks

Figure 4-14 – Comparison of Human and Mouse CpG Rich Regions Reveals Differences In Regulation of DNA Methylation (A) Statistics of all human and mouse CGRs. DNA methylation is in ESCs (B) Percent of all human and mouse CGRs with low, medium, or high CpG density (C) Percent of all human and mouse CGRs with low, medium or high DNA methylation, or in (D) CGRs with CpG density >0.75 only. (E). (F) For CGRs present in both species, percent of CGRs in each DNA methylation class, for human promoter CGRs with high DNA methylation, the corresponding mouse orthologue CGRs, and the corresponding mouse orthologues with CpG density > 0.75 .

Figure 4-15- Human and Mouse Promoter CGR DNA Methylation is Highly Similar For CGRs present in both species, percent of human and corresponding mouse promoter CGRs in each DNA methylation class

Table 1 : General Statistics for Human CGRs

Total Regions	173,308
Total bp Coverage	81,475,167
Total % Genome	2.5
Total % Non-Repeat Genome	2.7
Maximum CpG Number	1215
Maximum CpG Density	2.4
Maximum Size (bp)	16623
Maximum GC%	90
Average CpG Number	24.46
Average CpG Density	0.69
Average Size (bp)	470
Average GC%	58

Figure 4-1 - Nucleotide Properties of CpG Rich Regions

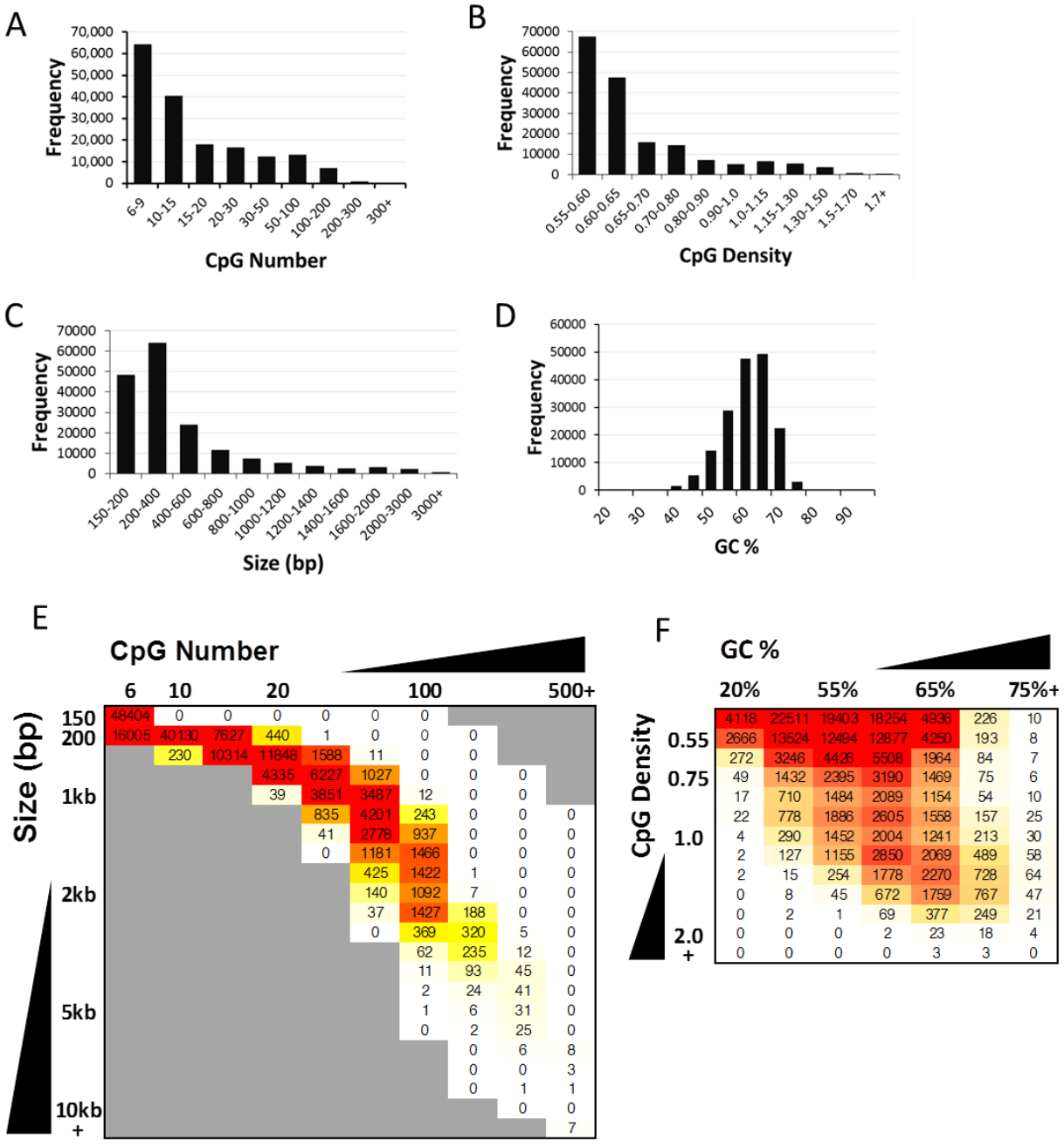


Figure 4-2 : CGR Location by Chromosome

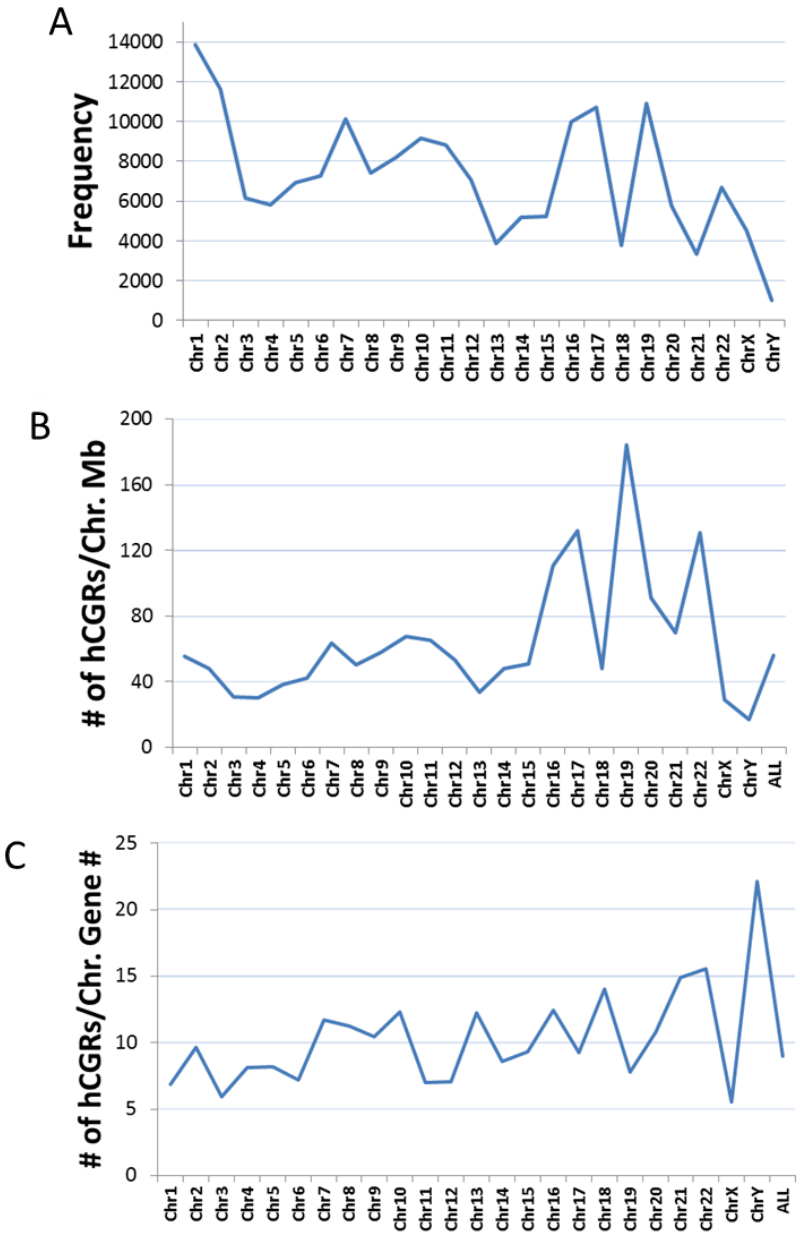


Figure 4-3 – Promoters are Enriched for CpG Dense Regions

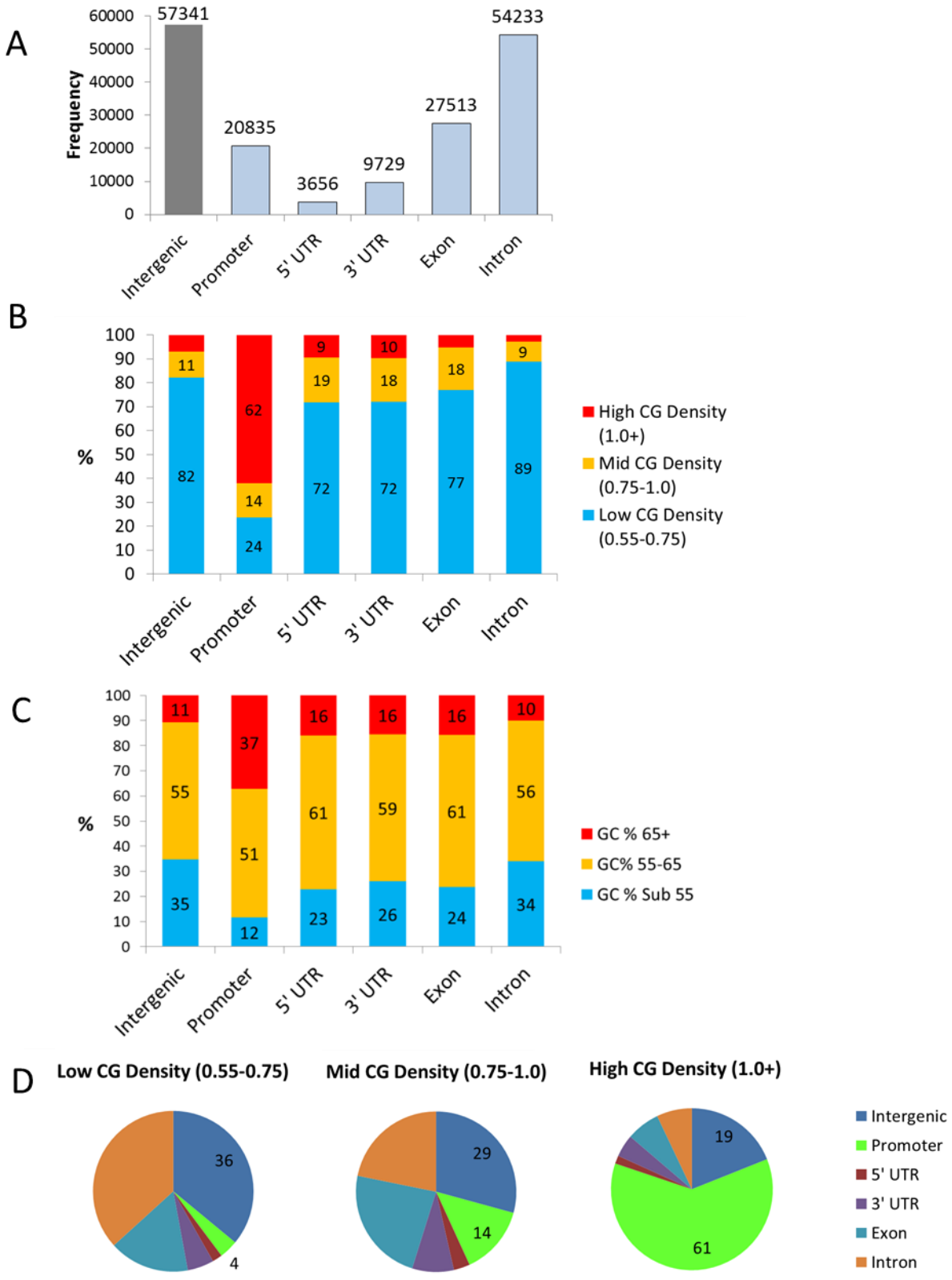


Figure 4-4 - Promoters are Enriched for High CG Number and Large CGR Size

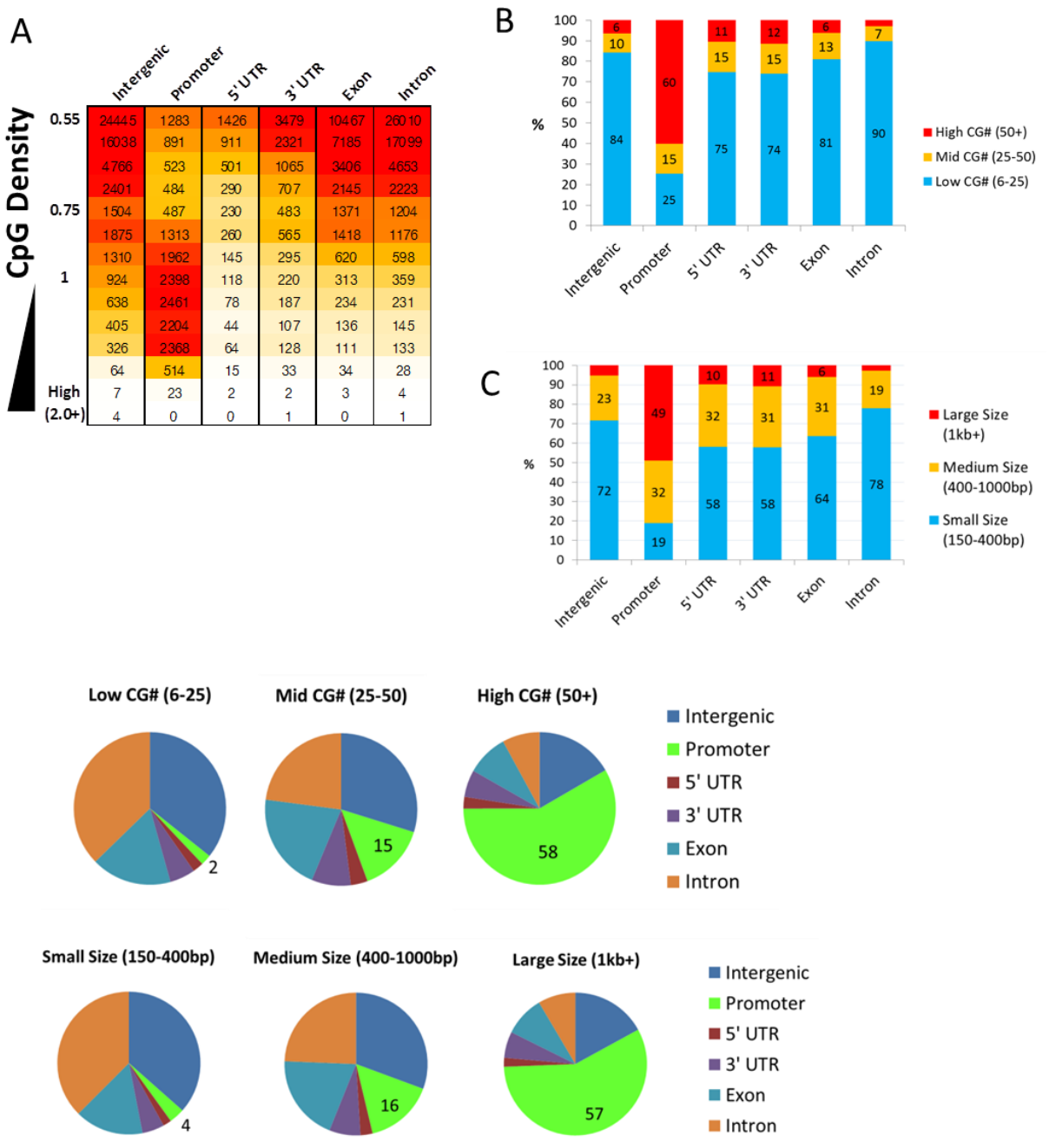
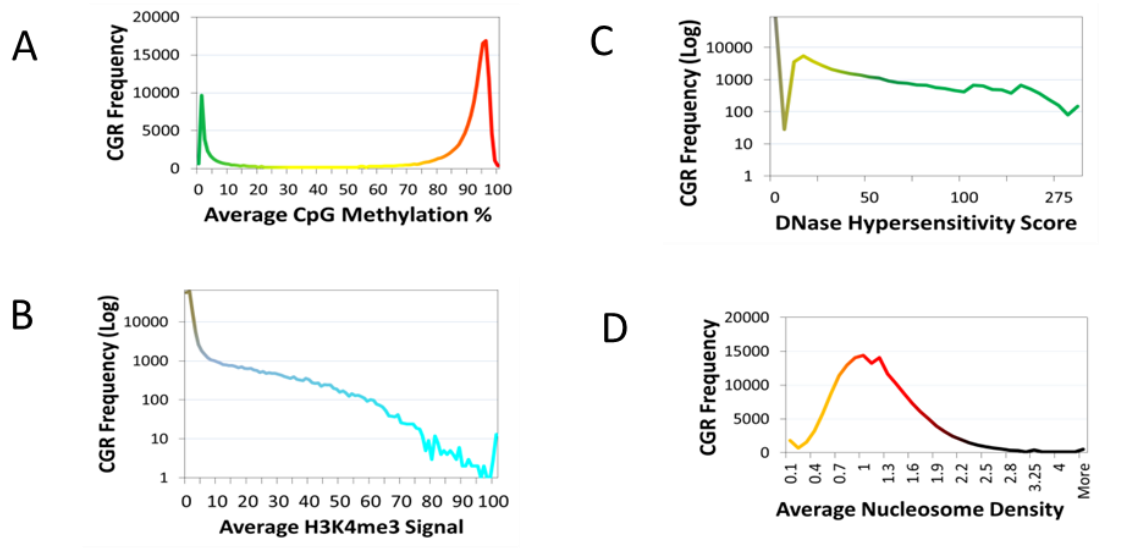


Figure 4-5 – Description of Chromatin Environment and Correlations at CpG Rich Regions



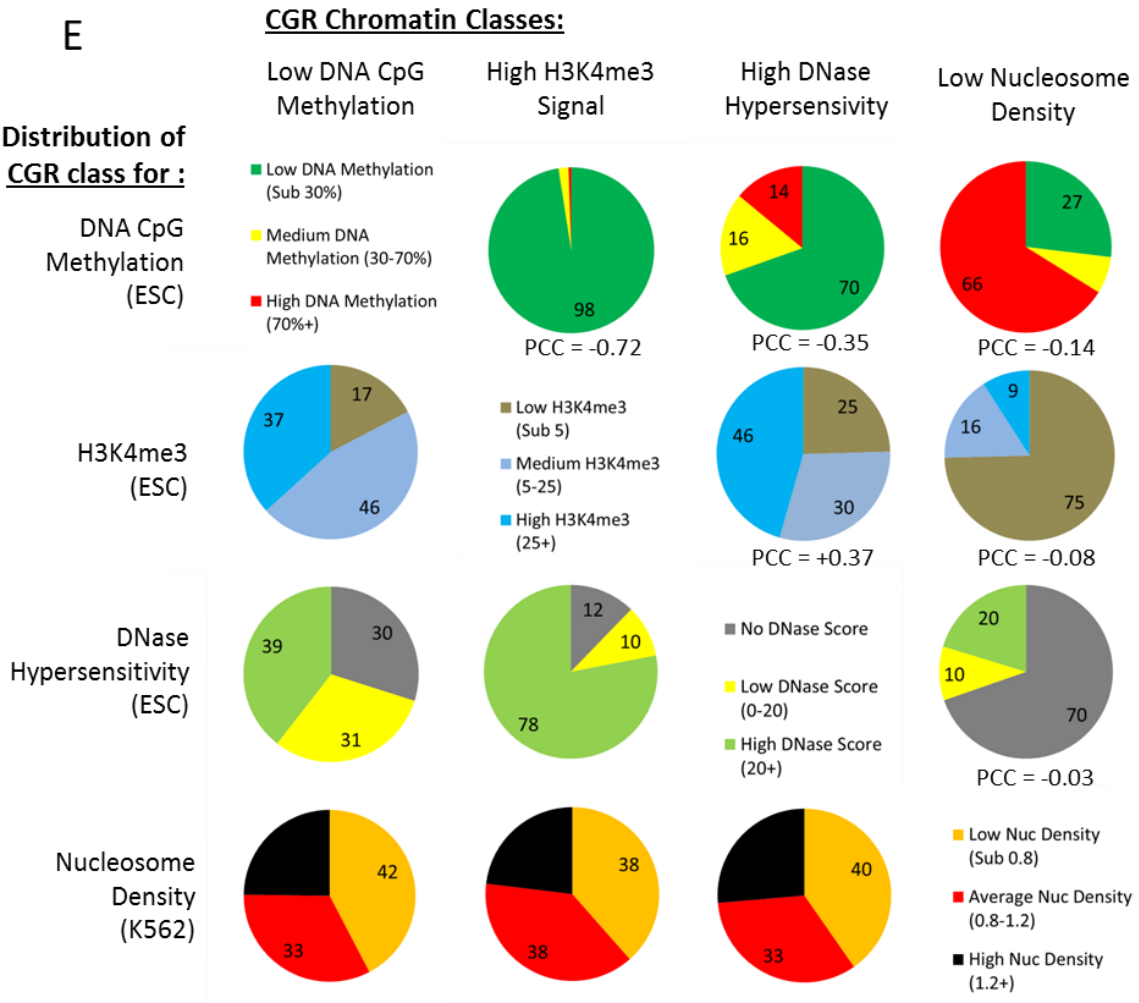


Figure 4-6 – CGRs with CpG Island Properties

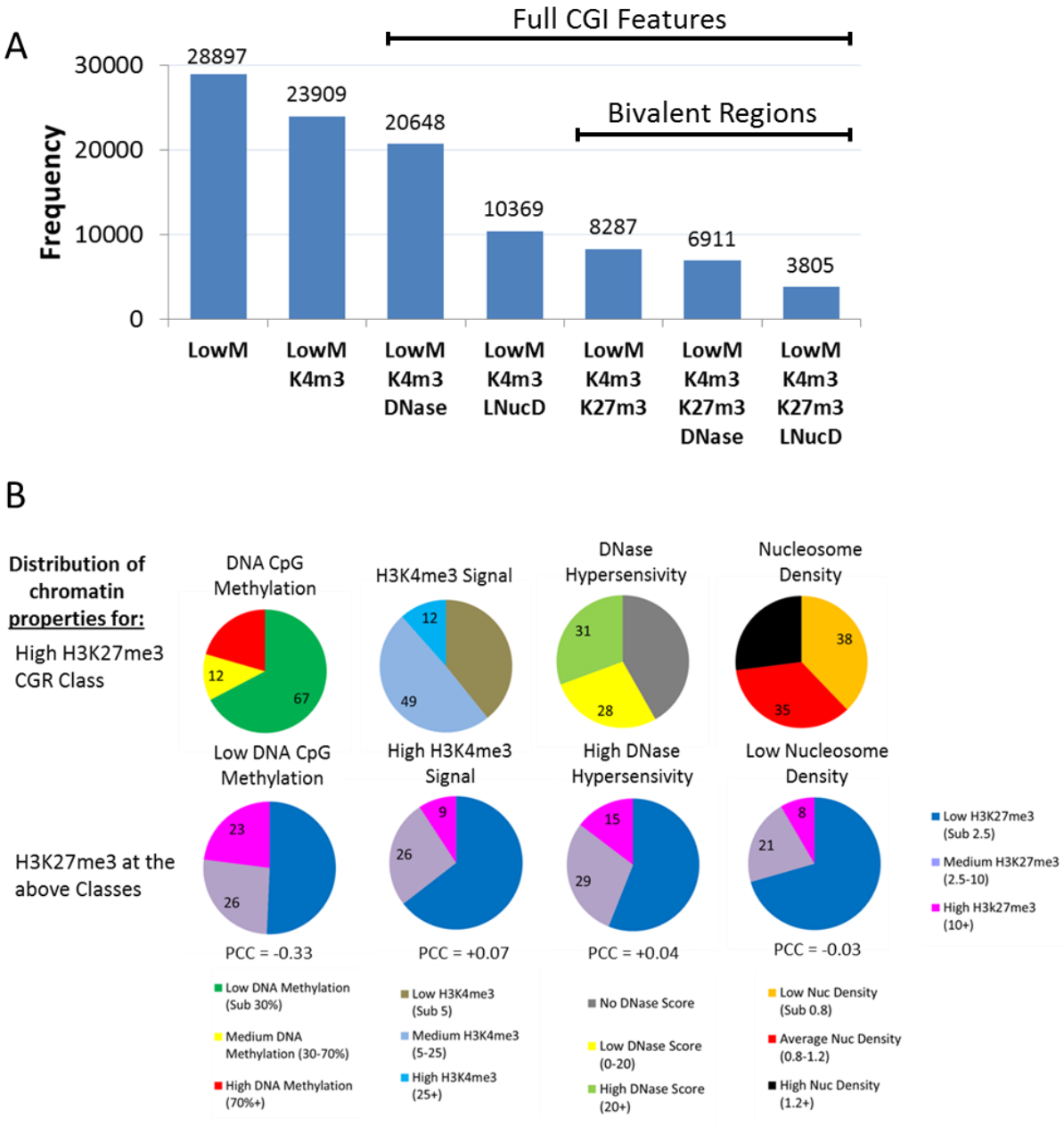


Figure 4-7 – Promoters CGRs are Enriched for CpG Island Features

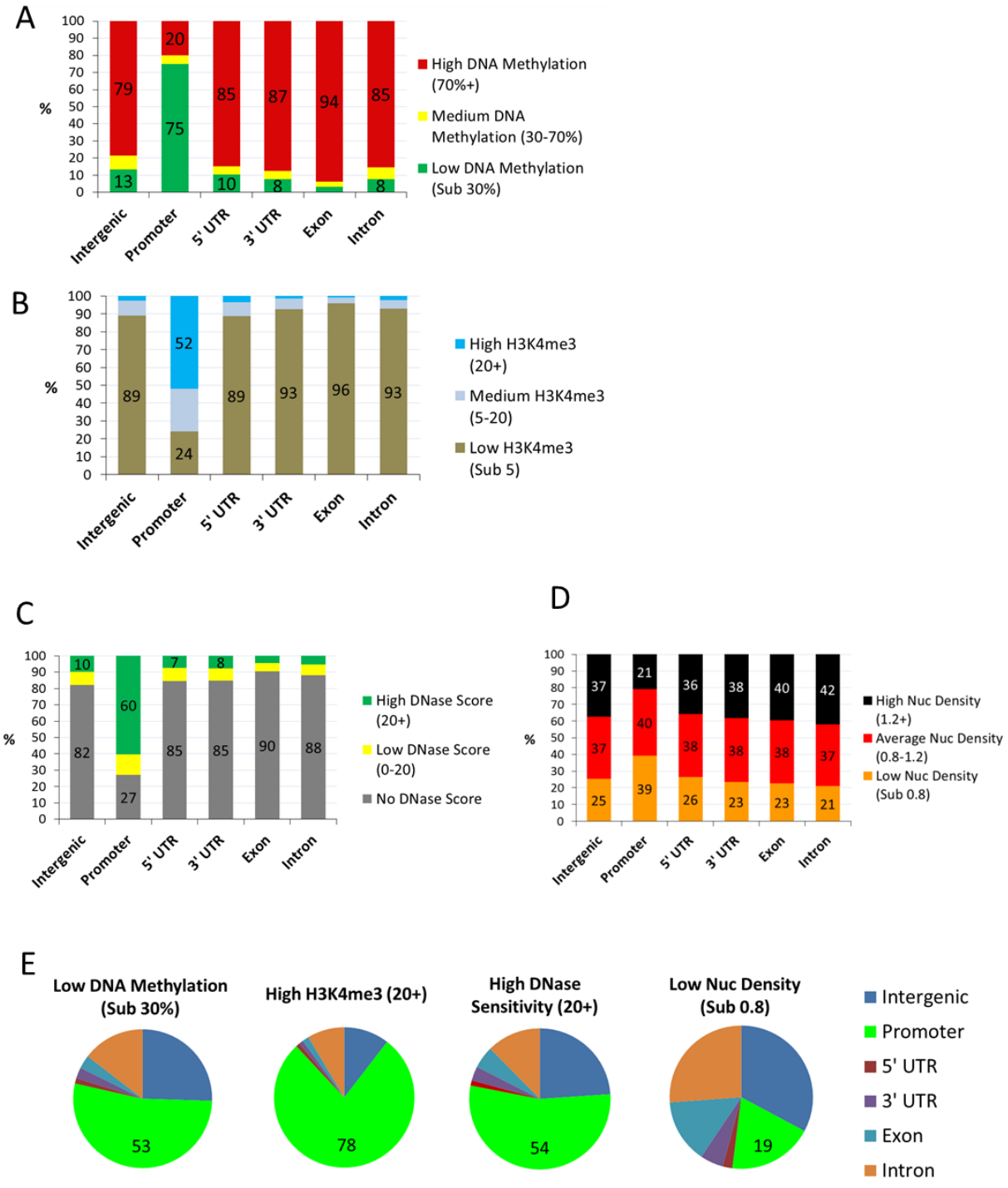
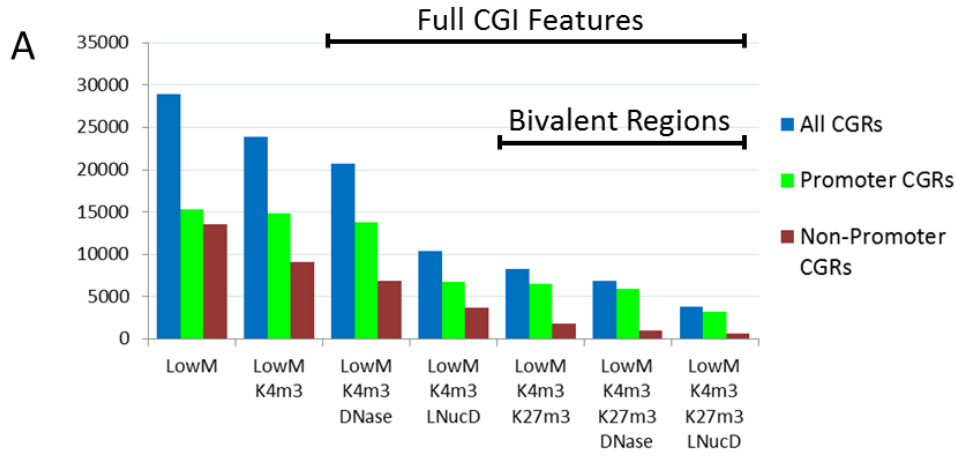
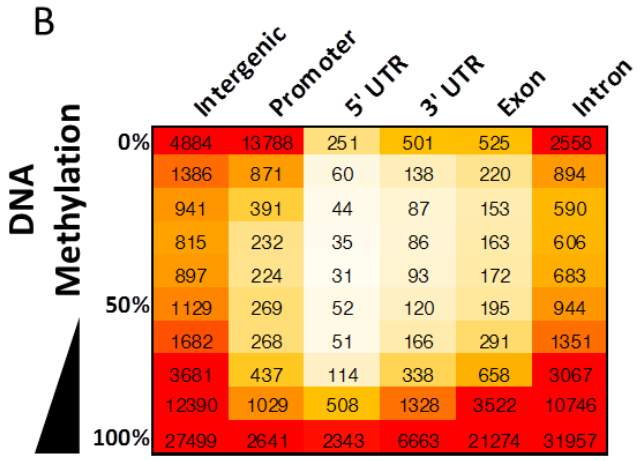
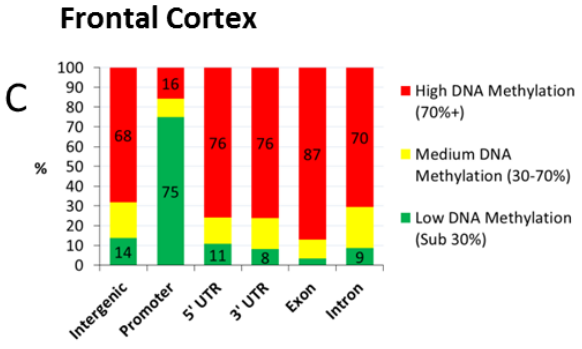
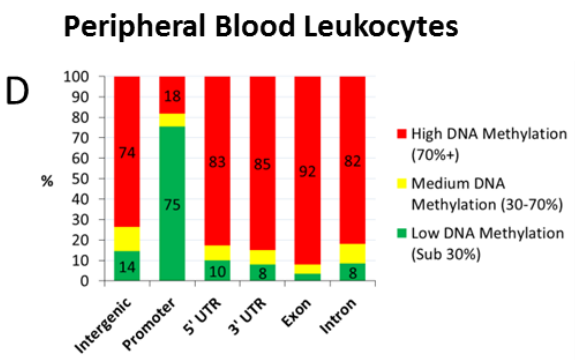


Figure 4-8 – DNA Methylation at Promoters, in Different Cell Types, and Other Chromatin Modifications









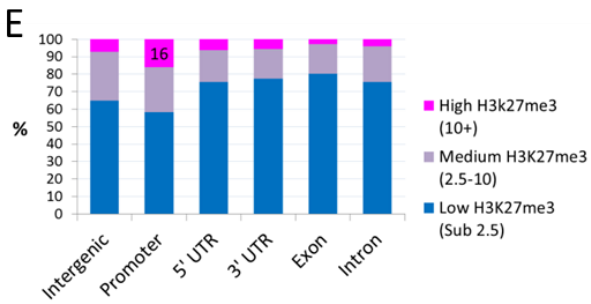


Figure 4-9 – Effect of Nucleotide Composition and Genomic Location on Chromatin at CpG Rich Regions

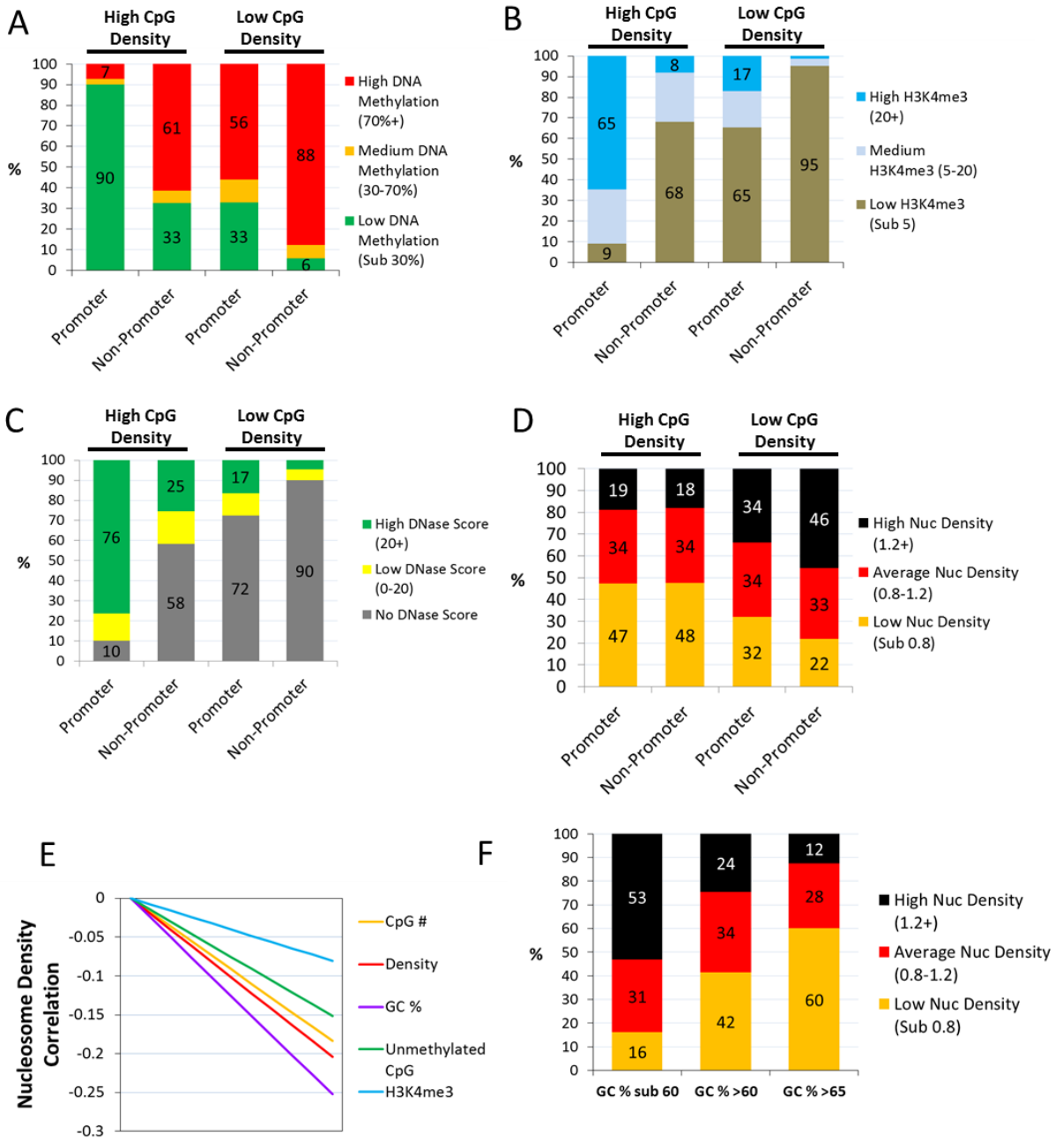
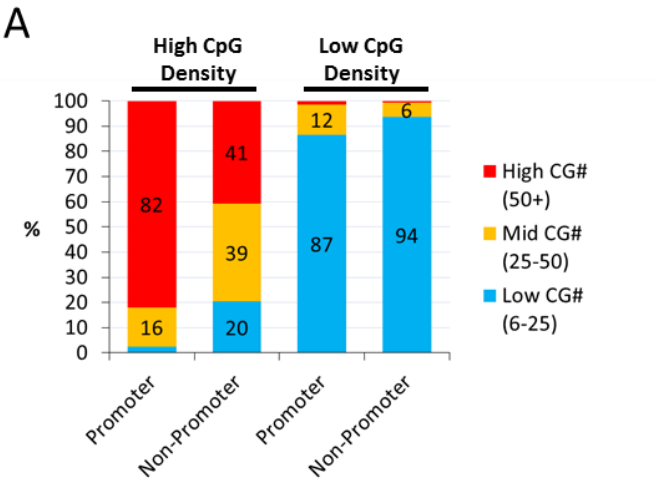
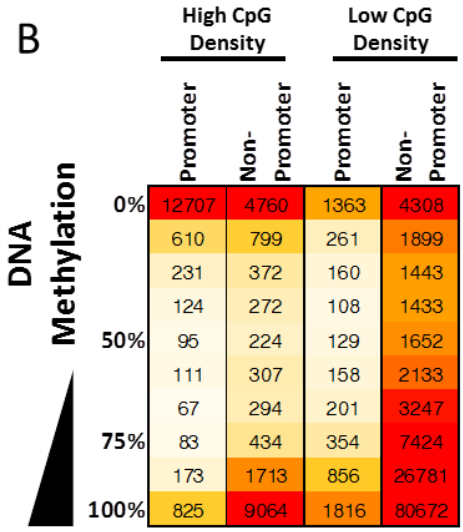
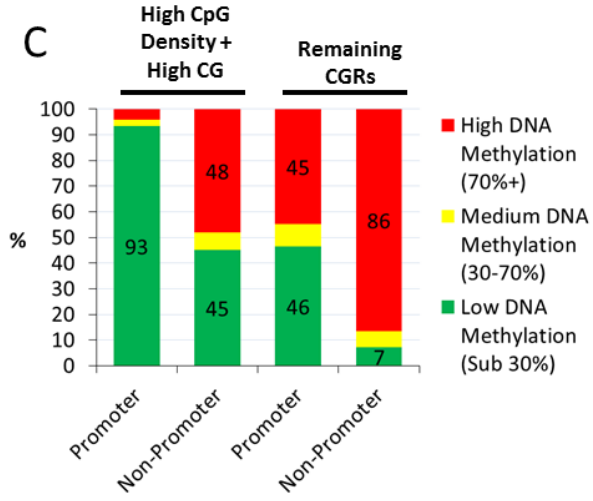


Figure 4-10 – Effect of Higher CpG Count and Density Cutoffs on Chromatin at CpG Rich Regions







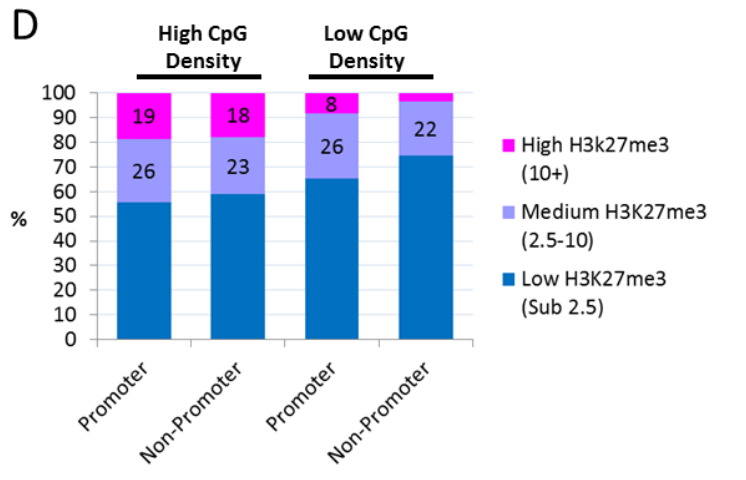


Figure 4-11 – Transcription factor binding enrichment at promoters affects chromatin features

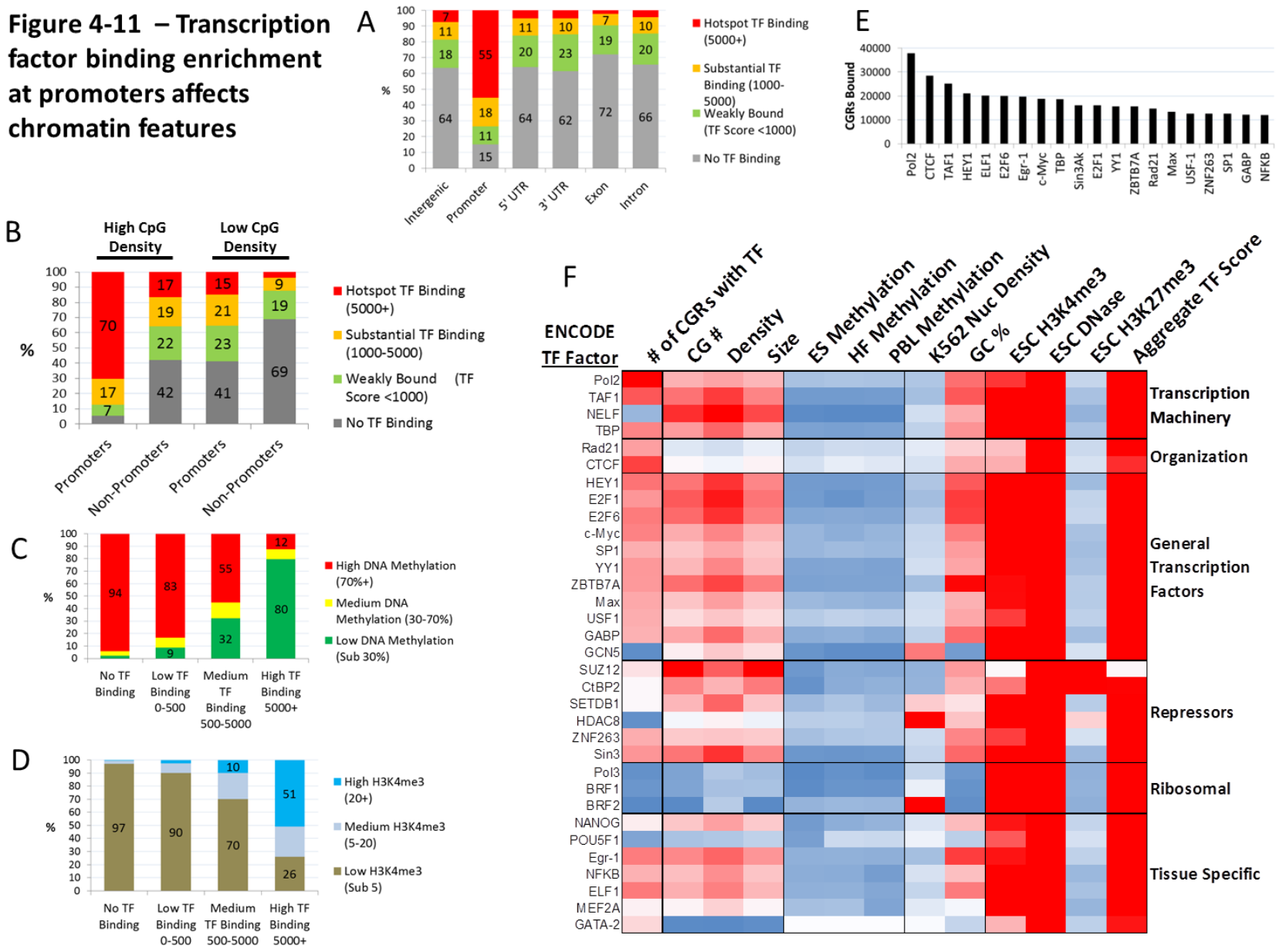


Figure 4-12 – Transcription Factor Binding at CGRs in Human ES Cells

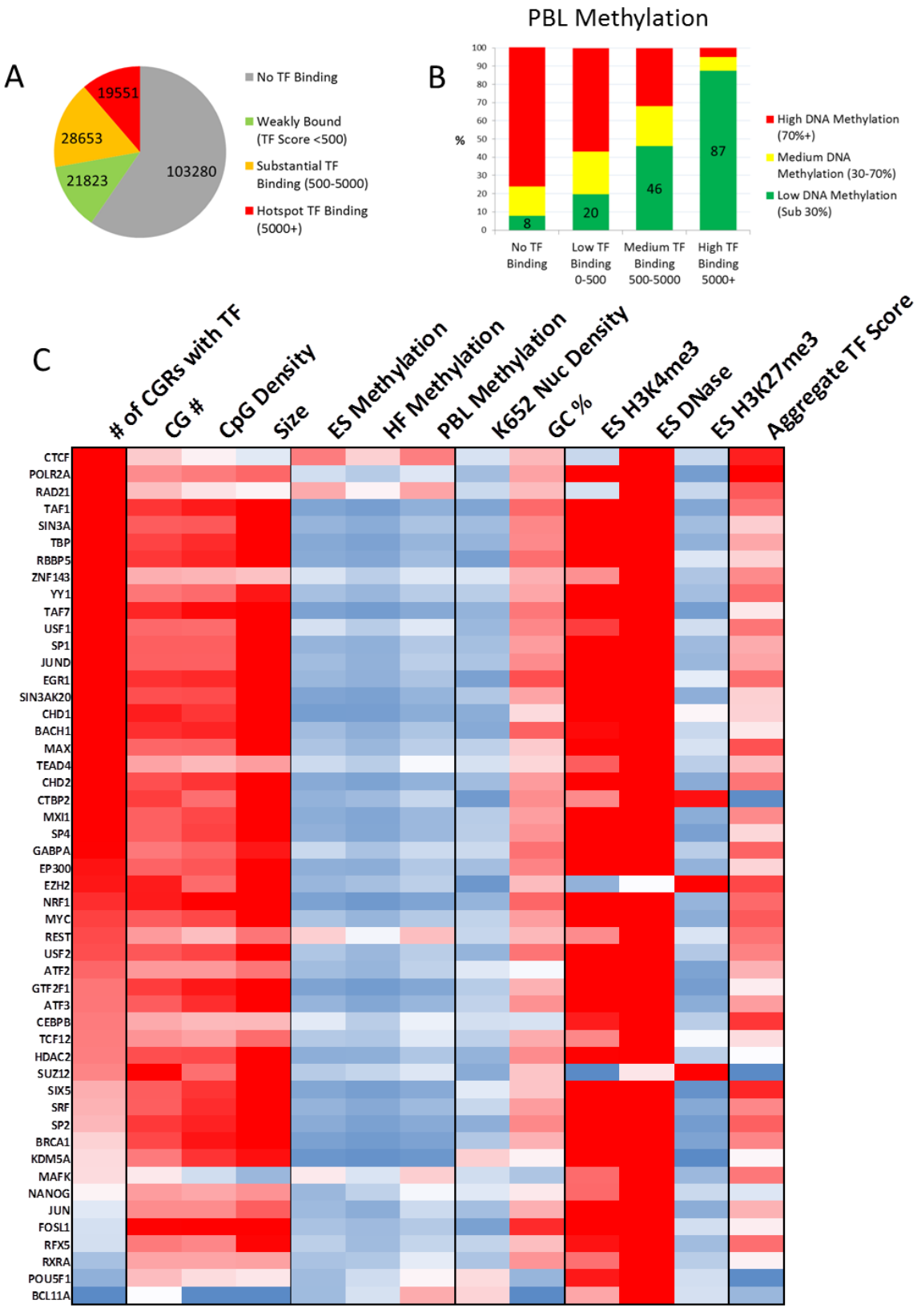


Figure 4-13 – Evidence of TF Binding Effects on CGR Chromatin

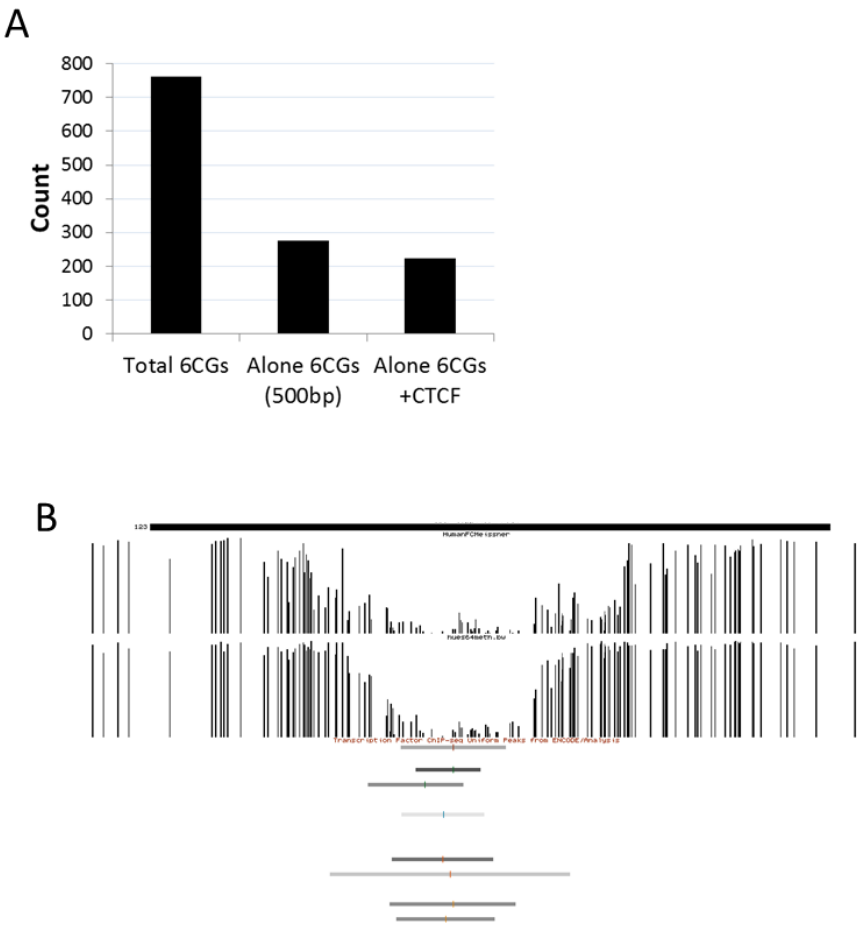
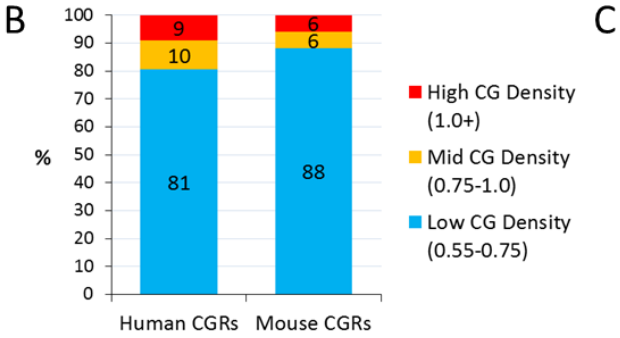
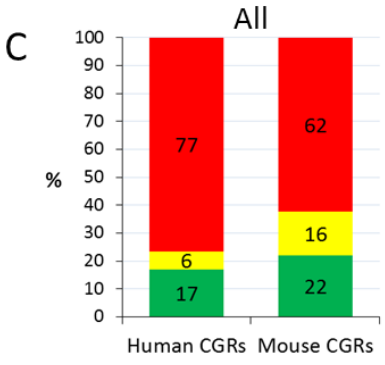


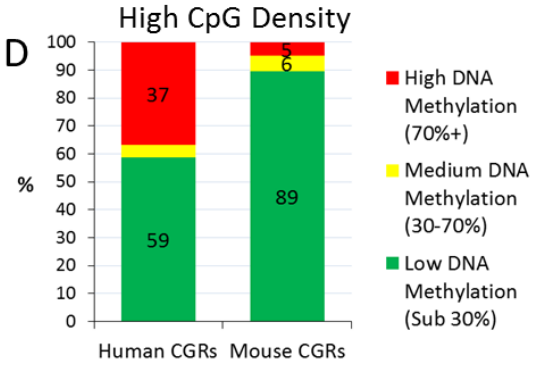
Figure 4-14 – Comparison of Human and Mouse CpG Rich Regions Reveals Differences In Regulation of DNA Methylation

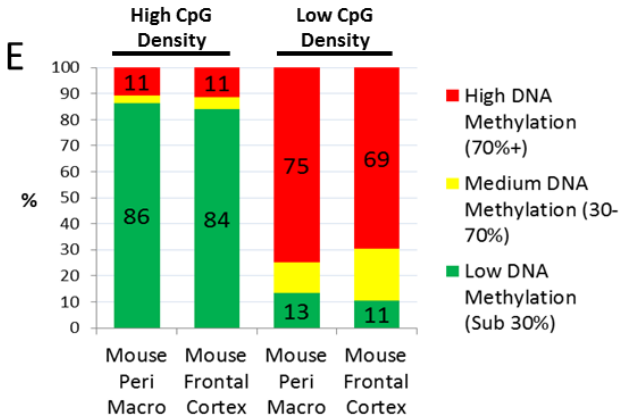
A

		Human	Mouse
	Total CG Rich Regions	173307	156908
Averages	CG#	24	17
	Density	0.69	0.66
	Size	470	356
	DNA Methylation %	74	63









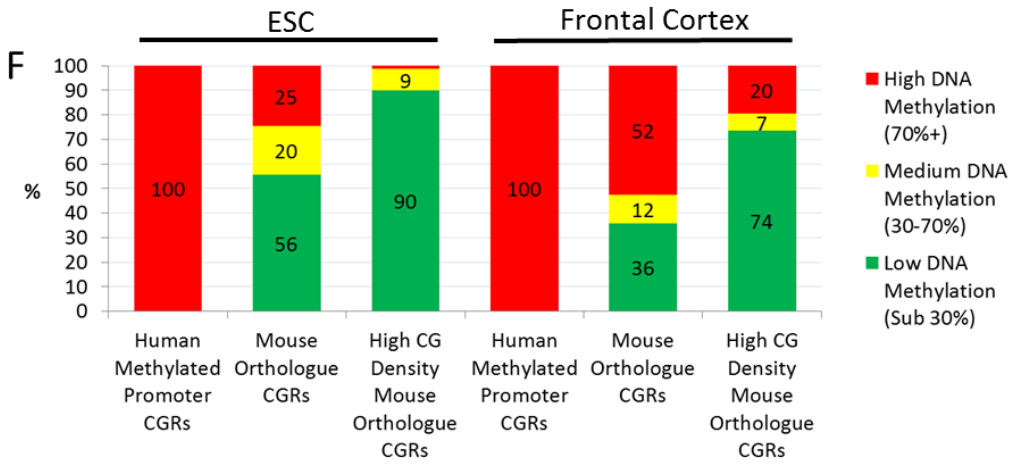
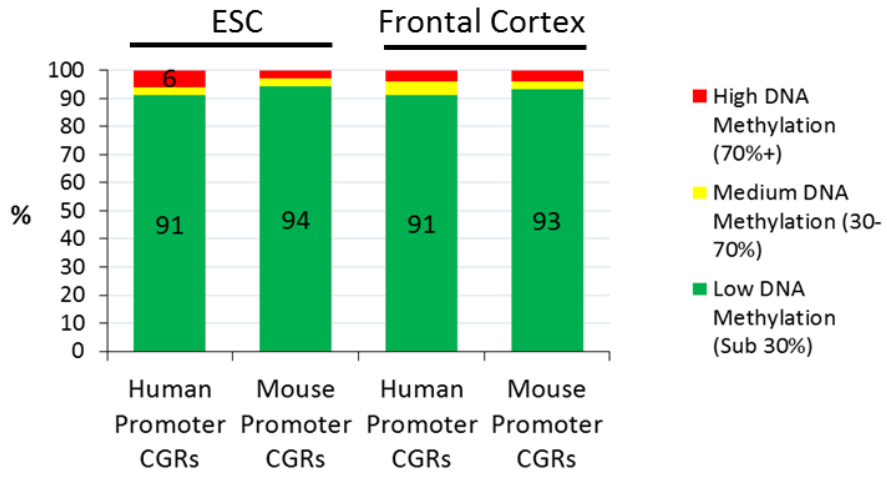


Figure 4-15 - Human and Mouse Promoter CGR DNA Methylation is Highly Similar



References

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bhatt, D.M., Pandya-Jones, A., Tong, A.-J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279–290.

Bird, A. (1985). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG Island Mapping by Epigenome Prediction. PLoS Comput. Biol. *3*, e110.

Chae, H., Park, J., Lee, S.-W., Nephew, K.P., and Kim, S. (2013). Comparative analysis using K-mer and K-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes. Nucleic Acids Res. *41*, 4783–4791.

Davuluri, R. V, Grosse, I., and Zhang, M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. *29*, 412–417.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fan, S., Zhang, M.Q., and Zhang, X. (2008). Histone methylation marks play important roles in predicting the methylation status of CpG islands. Biochem. Biophys. Res. Commun. *374*, 559–564.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., et al. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. *22*, 2399–2408.

Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell Stem Cell *2*, 160–169.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Holler, M., Westin, G., Jiricny, J., and Schaffner, W. (1988). Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. Genes Dev. *2*, 1127–1135.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet. *6*, e1001134.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, a. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K. a, Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V. V, and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. *31*, 89–97.

Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N.J., Harris, R. a., Xu, M., Chen, R., Shen, L., Milosavljevic, a., and Waterland, R. a. (2014). Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res. *42*, e43–e43.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Mukamel, E. a, Nery, J.R., Urich, M., Puddifoot, C. a, Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science (80-.). *341*, 1237905.

Macleod, D., Charlton, J., Mullins, J., and Bird, a P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282–2292.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P.J., Brown, D.T., and Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Dev. Cell *10*, 105–116.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Santos-Rosa, H., Schneider, R., and Bannister, A. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407–411.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. *103*, 1412–1417.

Shen, J., III, W.R., and Jones, P. (1992). High frequency mutagenesis by a DNA methyltransferase. Cell *71*, 1073–1080.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. *9*, 465–476.

Takai, D., and Jones, P. a (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. U. S. A. *99*, 3740–3745.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Vincent, J.J., Huang, Y., Chen, P.-Y., Feng, S., Calvopiña, J.H., Nee, K., Lee, S. a, Le, T., Yoon, A.J., Faull, K., et al. (2013). Stage-specific roles for tet1 and tet2 in DNA demethylation in primordial germ cells. Cell Stem Cell *12*, 470–478.

Xu, J., Watts, J. a, Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev. *23*, 2824–2838.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D. a, Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.

Chapter 5

Conclusion

Concluding Remarks

The discovery of CpG islands, due to their low methylation status and consequent vulnerability to the HP1 restriction enzyme, marks the beginning of research into unmethylated CpGs as a signal rather than a neutral basal state in mammals (Bird, 1985). It took many years for research to find unmethylated binding partner equivalents of proteins like MBD and MeCP2, but now it is known that DNA methylation is the genomic norm and unmethylated windows signal areas of special regulation. A growing body of evidence demonstrates these areas of special regulation are nearly always a consequence of direct transcription factor binding. The study of chromatin, at least at enhancers and CpG islands, is also necessarily a digression into transcription factor networks. However, transcription factor binding is oblique and ephermal while chromatin modifications can last a lifetime. The work presented herein is an attempt to define part of the mechanisms that drive chromatin modification by transcription factors at enhancers and CpG islands.

We have demonstrated more evidence for activity at tissue-specific enhancers during pluripotency, adding complexity to previous research (Xu et al., 2007, 2009). A broad stretch of the *Il12b* enhancer is capable of triggering a low methylation window. It is likely that many binding sites contribute to protection of the locus from DNA methylation, as targeted deletions were unable to block the low methylation window from forming. We did see slight increases in the methylation status at the *Il12b* enhancers with large deletions, suggesting that the multiple factors which bind here in ES cells function cooperatively. We also demonstrated that changes to the methylation status at the *Il12b* enhancer is strongly related to pluripotency and differentiation. Pluripotent cells have the highest level of methylation at the *Il12b* enhancer of any stable cell type, although it remains noticably below background. Upon differentiation to

embryoid bodies, we find that the methylation increases in aggregate up to genomic background levels, although the cells which successfully differentiate seem to remain below background. In differentiated macrophages, moderate *Il12b* enhancer methylation did not block *Il12b* expression and did not seem to correlate with the level of expression. However, each of the pluripotent lines tested in this experiment likely has a myriad of small differences which could take precedence over slight methylation at the *Il12b* enhancer. Although the link between establishment of the low methylation window in ES cells and later expression remains unclear, we have demonstrated a situation where transcription factors can act on the DNA methylation at a locus with variable penetrance determined by the cellular environment and developmental signals.

Tissue specific enhancers are necessarily affected by cell lineage changes, but the unmethylated DNA phenomena at CpG islands is usually constant across cell types (Suzuki and Bird, 2008). As such, CpG islands provide a powerful model system for studying the effect of underlying DNA content on chromatin. We studied the relationship between nucleotide content and the CpG island features of low DNA methylation, high H3K4me3, and low nucleosome occupancy and found that each feature has unique requirements. We characterized a small CpG island and found that a strong transcription binding site could result in low methylation locally and function to position nucleosomes adjacently, but did not drive H3K4me3. We also characterized the 601 positioning sequence, a CpG rich piece of DNA with the greatest known nucleosome binding affinity *in vitro* (Lowary and Widom, 1998). Despite the fact that 601 does not possess an activatory site, DNA methylation can spread through the region from adjacent transcription factor binding sites. Additionally we found that CpG rich DNA without activatory binding sites like 601 could still acquire low DNA methylation when they passed a certain size threshold *in vivo*. Our work is in agreement with several proposed mechanisms in the field,

supporting the importance of transcription factor binding for low methylation, the necessity for unmethylated CpGs to recruit H3K4me3, and the tendency for CpG rich DNA without strong trancription factor binding sites to acquire H3K27me3 (Lienert et al., 2011; Mendenhall et al., 2010; Thomson et al., 2010).

The experimental approach allowed us fine control over a limited selection of CpG rich sequences, so we increased the scope of our analysis by considering every CpG rich region genome wide. Computational derivation of CpG content and chromatin property correlations was made possible by the relatively recent exponential increase in publically available datasets, most notably including the ENCODE consortium (Bernstein et al., 2012). Our genome wide study found roughly 30,000 regions that suggest CpG content is stongly correlated with euchromatin features. The major predictive factor of the intensity of CpG island features at these regions was genomic location, specifically proximity to promoters of annotated genes. Expectedly, promoters are frequently bound by transcription factors, but surprisingly an extremely high amount of transcription factor binding was unique to promoters and strongly correlated with CpG island feature establishment. We observe that many of these transcription factors are a part of the Polymerase II machinery, in agreement with an experimental analysis of unmethylated CpG rich regions (Illingworth et al., 2010). Some of the frequent CpG island binding proteins we describe have already been correlated with demethylation function, including by our own experimental work for CTCF (Macleod et al., 1994; Stadler et al., 2011).

The only feature which seems to correlate with nucleotide content and is unaffected by other CpG island features is nucleosome occupancy. We demonstrated that the strongest determinant of low nucleosome occupancy in our genome wide studies is high GC content. Note that perterbations and positioning effects caused by CTCF, seen in our experiments and

described in more detail by others (Fu et al., 2008), do not cause low occupancy over the entire CpG island because it creates high density immediately adjacent. This does not preclude the involvement of transcription factors, but the level of trancription factor binding at CpG islands does not correlate at all to nucleosome occupancy. There is evidence suggesting GC nucleotide content intrinsically destabilizes nucleosome assembly but it is unknown how this applies *in vivo* (Ramirez-Carrozzi et al., 2009).The mechanism driving low nucleosome occupancy at some CpG islands in response to GC content requires more investigation.

Another issue that remains unresolved is the nature of the different between human and mouse CpG island regulation. In mouse cells, we have demonstrated both experimentally and with bioinformatics the existence of a CpG island size and density threshold triggering protection from DNA methylation. Mouse cells also share less than half of DNA methylation targeted CpG island promoters with human cells in both ES cells and frontal cortex. The CpG islands where DNA methylation is not conserved tend to be highly CpG dense in mice. The differential regulation of CpG island chromatin in response to underlying nucleotide content may result in important phenotypic differences between humans and mice. Delving further into the mouse dataset to discover which other chromatin properties behave similarly and to find other contributing nucleotide properties may reveal the source of this species specific difference.

Finally, we have described the important of promoters and TF binding for CpG features, but have not yet related these phenomena to transcription. Although there have been studies relating CpG content to transciption at both promoter CpG islands and distal CpG islands (Illingworth et al., 2010; Ramirez-Carrozzi et al., 2009), none have yet taken into account the nucleotide features of CpG islands at the detail our dataset can provide. RNA-sequencing datasets are being rapidly being produced from nearly every cell type and condition. Combined

with our knowledge about CpG island chromatin we may be able to accurately define the effect of CpG islands on transcription in future studies.

References

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bird, A. (1985). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet. *4*, e1000138.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet. *6*, e1001134.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097.

Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. *276*, 19–42.

Macleod, D., Charlton, J., Mullins, J., and Bird, a P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282–2292.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. *9*, 465–476.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J. a, Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG

dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *104*, 12377–12382.

Xu, J., Watts, J. a, Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev. *23*, 2824–2838.