UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Naples, Italy

DOCTORAL THESIS

Computational Biology and Bioinformatics

XXV cycle

Epigenetic modifications in CpG islands and signatures of selective pressure in human genome

Tutor: Prof. Sergio Cocozza

Co-tutor: Prof. Gennaro Miele

PhD candidate: Most. Mauluda Akhtar

March 2013

Duration of PhD: 3 years (March 2010 to February 2013)

To
My Parents
&
My Husband

The Nucleus of My Life.....!

Contents

Summary

Epigenetics deals the heritable changes in gene regulation which is not related with the changes of DNA sequence itself. Among several molecular mechanisms that mediate epigenetic phenomena, DNA methylation and histone modifications are well known markers. CpG islands (CGIs) are the key epigenomic elements in mammalian genome. CGIs are defined as the segments of the genome that show increased level of CpG dinucleotides and GC content. These CGIs are enriched at genes, about 60% of all genes in the human genome containing a CGI upstream. DNA methylation at CGIs is one of the most intensively studied epigenetic mechanisms. It is fundamental for cellular differentiation and control of transcriptional potential. DNA methylation is involved also in several processes that are central to evolutionary biology, including phenotypic plasticity and evolvability. Furthermore, histone modifications in CGIs are associated with the changes in chromatin states and with transcription activity. Changes in gene expression play a crucial role in adaptation and evolution. Considering the role of DNA methylation and histone modifications in gene expression changes, our aim was to explore the relationship between these two epigenetic marks and selective pressure in human genome.

In the first step, we explored a relationship between CpG islands methylation and signatures of selective pressure in *Homo sapiens*, using a computational biology approach. For this we analyzed methylation data of 25 human cell lines from the Encyclopedia of DNA Elements (ENCODE) Consortium. To define regions under selective pressure, we used three distinct signatures that mark selective events from different evolutionary periods. We compared the DNA methylation of CpG islands in genomic regions under selective pressure with the methylation of CpG islands in the remaining part of the genome. We found that CpG islands in the regions under selective pressure are undermethylated than the CpG islands of the other group.

In the second step, we have studied, using a computational biology approach, the relationship between histone modifications in CGIs and selective pressure in *Homo sapiens*. We considered three histone modifications: histone H3 lysine 4 trimethylation (H3K4me3), acetylation of histone H3 at lysine 27 (H3K27ac) and trimethylation of histone H3 at lysine 36 (H3K36me3), and we used the publicly available genomic-scale histone modification data of 23 human cell lines. To define regions under selective pressure, we used the similar approach as used in the first step. We found that, CGIs under selective pressure showed significant enrichments for histone modifications.

In conclusion, our overall findings suggest that CpG islands that experienced selective pressure are characterized by distinct epigenetic signatures.

Chapter 1: Background

Epigenetics and its mediators

Epigenetics has been one of the most thrilling terms in the field of biological research. Literally the term "epigenetic" means "in addition to changes in genetic sequence." Conrad Waddington in 1942 first mentioned ''Epigenetics'' to interpret the process during development by which genotype gives rise to phenotype (Waddington 1942). Since the term was introduced to the scientific community, various researches suggested that gene function could be altered by more than just the changes in gene sequence. The modern and generally accepted definition of epigenetics is the study of changes in gene function that are stably heritable and that do not entail a change in DNA sequence (Berger et al. 2009). Over the last few years, the development of several genomic and proteomic technologies like next-generation sequencing (NGS), chromatin immuno precipitation (ChIP-Seq) have provided a broader view of the epigenome which refers to the complete description of these potentially heritable changes across the genome (Park 2009; Dunham et al. 2012).

Several molecular mechanisms mediate epigenetic phenomena including DNA methylation, histone modification, chromatin remodeling and micro RNA (Tammen et al. 2012). Among those, DNA methylation and histone modifications (table 1.1) are the most studied epigenetic mediators. The role of epigenetics in various diseases especially in cancer is well studied (Dawson and Kouzarides 2012). However, studies concerning the link between epigenetics and evolution are still limited (see the paragraph epigenetics and evolution).

Table 1. Chromatin Modifications, Readers, and Their Function

Chromatin Modification	Nomenclature	Chromatin-Reader Motif	Attributed Function
DNA Modifications			
5-methylcytosine	5mC	MBD domain	transcription
5-hydroxymethylcytosine	5hmC	unknown	transcription
5-formylcytosine	5fC	unknown	unknown
5-carboxylcytosine	5caC	unknown	unknown
Histone Modifications			
Acetylation	K-ac	BromodomainTandem, PHD fingers	transcription, repair, replication, and condensation
Methylation (lysine)	K-me1, K-me2, K-me3	Chromodomain, Tudor domain, MBT domain, PWWP domain, PHD fingers, WD40/β propeller	transcription and repair
Methylation (arginine)	R-me1, R-me2s, R-me2a	Tudor domain	transcription
Phosphorylation (serine and threonine)	S-ph, T-ph	14-3-3, BRCT	transcription, repair, and condensation
Phosphorylation (tyrosine)	Y-ph	SH2[a]	transcription and repair
Ubiquitylation	K-ub	UIM, IUIM	transcription and repair
Sumoylation	K-su	SIM[a]	transcription and repair
ADP ribosylation	E-ar	Macro domain, PBZ domain	transcription and repair
Deimination	R→Cit	unknown	transcription and decondensation
Proline isomerisation	P-cis⇔P-trans	unknown	transcription
Crotonylation	K-cr	unknown	transcription
Propionylation	K-pr	unknown	unknown
Butyrylation	K-bu	unknown	unknown
Formylation	K-fo	unknown	unknown
Hyroxylation	Y-oh	unknown	unknown
O-GlcNAcylation (serine and threonine)	S-GlcNAc; T-GlcNAc	unknown	transcription

Modifications: me1, monomethylation; me2, dimethylation; me3, trimethylation; me2s, symmetrical dimethylation; me2a, asymmetrical dimethylation; and Cit, citrulline. Reader domains: MBD, methyl-CpG-binding domain; PHD, plant homeodomain; MBT, malignant brain tumor domain; PWWP, proline-tryptophan-tryptophan-proline domain; BRCT, BRCA1 C terminus domain; UIM, ubiquitin interaction motif; IUIM, inverted ubiquitin interaction motif; SIM, sumo interaction motif; and PBZ, poly ADP-ribose binding zinc finger.

[a]These are established binding modules for the posttranslational modification; however, binding to modified histones has not been firmly established.

Table 1.1: Chromatin modifications, readers, and their functions (Dawson and Kouzarides 2012).

DNA methylation

In mammalian cells, DNA methylation is a chemical modification by a methyl (-CH3) group added by the enzymatic family of DNA methyltransferases (DNMTs) on C5 position of cytosine in DNA molecule that results 5-methylcytosine (Figure 1.1). It occurs almost exclusively at CpG dinucleotides where a cytosine nucleotide occurs next to a guanine nucleotide (Bird 2002; Jones and Liang 2009). In human somatic cells, ~70–80% of cytosine in CpG sites is methylated (Chen and Riggs 2011).

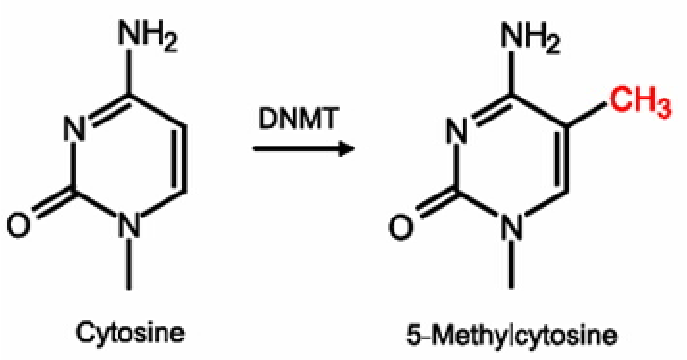

Figure 1.1: DNA methylation; addition of methyl group in 5C of cytosine

Although most CpG dinucleotides are methylated, the genome also contains CpG islands (CGIs), a short CpG-rich genomic regions, which are usually unmethylated (discussed below). Though methylation in other sequences has been reported in mammals (Lister et al. 2009), the function of non-CpG methylation is currently unknown. DNA methylation is involved in several key processes like X chromosome inactivation, genomic imprinting and silencing of germline specific genes and repetitive elements (Bird 2002; Jones and Liang 2009), and is essential for normal development (Eckhardt et al. 2006). Methylation at CGIs promoter is often associated with the shutdown of respective genes and aberrant DNA methylation is found to occur in various types of cancer, leading to silencing of some tumor suppressor genes (Bird 2002; Dawson and Kouzarides 2012; Lopez-Serra and Esteller 2012). In mammal, three different DNMTs namely DNMT1, DNMT3a, and DNMT3b catalyze and maintain DNA methylation throughout the cell cycle (Jones and Liang 2009). Among the three enzymes DNMT1, which has preference for hemimethylated CpG sites, is mainly involved to copy pre-existing methylation patterns to the newly synthesized strand during the cell cycle (Jones and Liang 2009) probably with the help of the protein UHRF1 which binds hemimethylated sites. In addition, both DNMT3a and DNMT3b are *de novo* methyltransferases and they have no affinity for hemimethylated CpG substrates *in vitro* (Okano et al. 1998; Gowher and Jeltsch 2001). DNMT3a/3b are reported to be responsible to establish methylation patterns during early development (Okano et al. 1999). *De novo* methylation by DNMT3a/3b also contributes to the maintenance of DNA methylation patterns (Liang et al. 2002; Chen et al. 2003), perhaps by methylating CpG sites overlooked by DNMT1 (Riggs and Xiong 2004; Jones and Liang 2009).

Histone modification

Chromatin is a complex of DNA and proteins (histones). Histone proteins help packaging the DNA by forming nucleosome around which 147 base pairs of DNA are wrapped. Nucleosome is a basic unit of DNA composed of an octamer consisting two copies of four core histones: H3, H4, H2A and H2B (Luger et al. 1997).

Histone proteins can be modified covalently by different molecules in their N-terminal tails (Table 1.1). Mainly studied covalent modification of histones includes acetylation of lysines, methylation of lysines and arginines, phosphorylations of serines and threonines (Figure 1.2).

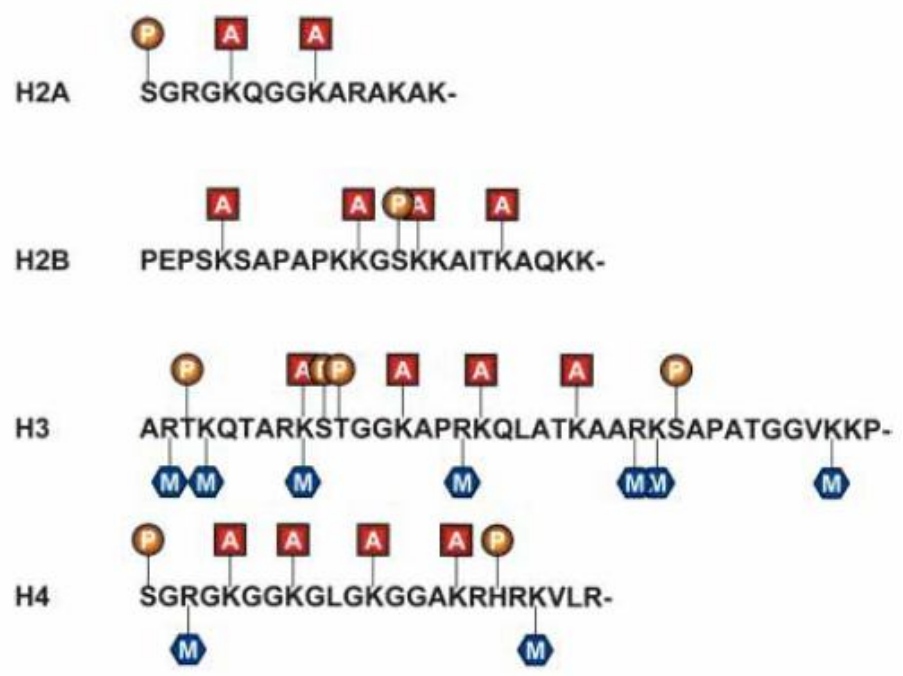

Figure 1.2. Covalent modifications of the N-terminal tail of the canonical core histones. Phosphorylations are shown as yellow circles, acetylations as red squares, and methylations as bluehexagons (Lund and Van Lohuizen 2004).

These modifications can either activate or repress transcription of associated genes depending on the types of amino acids and modifications (Kouzarides 2007; Blomen and Boonstra 2011). The best-characterized sites of histone modifications are those that occur on lysine residues. Methylation of some lysine residues (e.g. H3K4, H3K36 and H3K79) is often associated with active genes, whereas others (H3K9, H3K27 and H4K20) are associated with inactive genes. In general, acetylation of lysine is linked to transcriptional activation (Barski et al. 2007; Kouzarides 2007).

In our study, we are particularly interested on histone H3 lysine 4 trimethylation (H3K4me3), histone H3 lysine 27 acetylation (H3K27ac) and histone H3 lysine 36 trimethylation (H3K36me3) considering their role in gene activations.

***H3K4me3*:** H3K4me3 is functionally important histone mark, associated with transcription activation (Kouzarides 2007). It is found that H3K4me3 clearly peak at 5' ends of annotated human genes (Guenther et al. 2007). It is intriguing to find the co-occurrence of the activating H3K4me3 with the repressive H3K27me3 mark in embryonic stem cells (ESCs) on silent transcription state (Bernstein et al. 2006; Mikkelsen et al. 2007). These bivalent domains are poised between the states, active transcription or stable repression. Upon differentiation, losing H3K27me3, they can become active or be subject to more stable transcriptional repression (Deaton and Bird 2011). Like DNA methylation, H3K4me3 mark plays a crucial role in mammalian development (Bernstein et al. 2006). Alteration of this mark is found to be associated with cancer and other diseases (Kaneda et al. 2009; Ke et al. 2009;

Sandgren et al. 2010). In mammals, the trimethylation of H3K4 is maintained by different histone methyltransferases, such as MLL1 or ASH1L (Dou et al. 2006; Gregory et al. 2007). The discovery of histone lysine demethylase enzymes indicates that histone methylation is a biochemically dynamic state (Shi et al. 2004). Lysine demethylase enzymes for H3K4me3 have been found on active promoters indicating that methylation of H3K4me3 on active genes may display a cyclic behavior. KDM5B/Jarid1B/PLU1 belonging to the JmjC domain-containing family of histone demethylases, is an H3K4me3/me2-specific lysine demethylase (Lloret-Llinares et al. 2012)

***H3K27ac*:** The acetylation of lysine residues is a major histone modification involved in transcription, chromatin structure, and DNA repair (Dawson and Kouzarides 2012). Among all lysine acetylations H3K27ac is known as promoter mark associated with transcriptional activation (Wang et al. 2008). H3K27ac is a evolutionarily conserved mark among species (Woo and Li 2012). Also this mark is an important enhancer mark that can distinguish active from poised enhancer elements (Creyghton et al. 2010). Acetylation of H3K27 is catalyzed by the acetyltransferases p300 and CBP (Tie et al. 2009; Pasini et al. 2010). p300 and CBP are important transcriptional and epigenetic regulators (Kalkhoven 2004) and dysregulation of their functions is associated with leukemia and other types of cancers. Consequently, p300 and CBP are generally considered as anti-cancer drug target (Wang et al. 2013). The NuRD (nucleosome remodelling and deacetylation complex), which is required for lineage commitment of pluripotent cells, mediates deacetylation of histone H3K27 and recruits Polycomb Repressive Complex 2 (PRC2) that in turns is necessary for subsequent H3K27 trimethylation at NuRD target promoters. It seems that NuRD controls the balance between acetylation and methylation of histones, thus precisely directs the genes expression crucial for embryonic development (Reynolds et al. 2011).

H3K36me3: H3K36me3 is a gene body mark (Barski et al. 2007), evolutionary conserved between human and mouse (Kolasinska-Zwierz et al. 2009). This mark generally is associated with gene activation (Wang et al. 2008) specifically transcriptional elongation (Li et al. 2007). H3K36me3 is also found to be associated with alternative splicing, transcriptional repression, dosage compensation, DNA replication and repair, DNA methylation and the transmission of the memory of gene expression from parents to offspring during development (Wagner and Carpenter 2012). In mammalian cells, histone methyltransferase SETD2 is thought to be the only trimethylase of H3K36 (Edmunds et al. 2007). Loss of SETD2 is found to be involved in the development of sporadic clear renal cell carcinoma (Duns et al. 2010). SETD2 has also been hypothesized to be a tumour suppressor in breast cancer (Newbold and Mokbel 2010). Histone demethylase JHDM3A (jumonji C (JmjC)-domain-containing histone demethylase 3A; also known as JMJD2A) is capable of removing trimethylation of H3K36 (Klose et al. 2006).

CpG island (CGIs), a platform for epigenetic gene regulation

CGIs are considered as the key epigenomic elements in mammalian genome. CGIs were first defined as the fraction of genome characterized by low level of DNA methylation (Bird et al. 1985). Later, based on the CpG content of the DNA sequence, CGIs were defined as the segments of the genome that show increased level of CpG dinucleotides and GC content (Gardiner-Garden and Frommer 1987; Takai and Jones 2002). CGIs are considered to be the promoter mark of vertebrate genome since approximately 70% of annotated gene promoters are associated with a CGI (Saxonov et al. 2006). This association suggests the regulatory role of CGIs. It is well established that CpG sites in promoter CGIs

are undermethylated in expressed genes, while hypermethylation of promoter CpG sites is associated with gene silencing (Jones 2012). Almost all the housekeeping genes, some tissue-specific genes and developmental regulator genes are associated with CGIs (Larsen et al. 1992; Zhu et al. 2008).

CGIs that do not co-localize with annotated promoters have been found in intragenic, 3', and intergenic regions (Medvedeva et al. 2010). These kinds of CGIs were termed as ''orphan'' CGIs in another study (Illingworth et al. 2010). Though the functions of orphan CGIs is not well understood yet, it has been reported that some orphan CGIs might represent alternative promoters of nearby annotated genes (Maunakea et al. 2010). A subset of intergenic orphan CGIs having peaks of H3K4me3 are assocoated with transcription start sites (TSSs) for long noncoding RNAs (Guttman et al. 2009).

Using evolutionary modelling, Cohen et al. reclassified CGIs as hypo deaminated CGIs, ~80% of these are present within 10 kb of an TSS and show strong overlap with H3K4me3; and those that had arisen as a by-product of biased gene conversion (BGC), are typically constitutively hypermethylated (Cohen et al. 2011). Recently, another study identified non-polymorphic species-specific CpG dinucleotides (termed "CpG beacons") as a distinct genomic feature associated with CGI evolution, human traits and disease (Bell et al. 2012).

CGIs can either be permissive or repressive for transcription mediated by epigenetic state. By building a chromatin-based atmosphere CGIs are thought to contribute to the functional output of related genes (Blackledge and Klose 2011).

Permissive state

Though CGIs are associated with mammalian gene promoters but still it is not completely understood how CGIs execute their role in gene regulation. Earlier studies reported that high levels of histone H3 and H4 acetylation which are the marks of active chromatin, are associated with CGIs (Tazi and Bird 1990; Birney et al. 2007; Wang et al. 2008; Su et al. 2010). There is a robust correlation between the active promoter mark, H3K4me3 sites and CGIs (Clouaire et al. 2012). CpG density in CGIs positively correlates to H3K4me3 level mediated by Cfp1 (CxxC finger protein 1), that recognizes non-methylated CpG is a fundamental component of the Setd1 H3K4 methyltransferase complex (Voo et al. 2000; Lee and Skalnik 2005; Illingworth et al. 2010). Several transcription factors like Sp1, CREB (Mancini et al. 1999) and CTCF (Renda et al. 2007) contain CpG in their binding site but CpG methylation often blocks their recognition sites.

Histone mark H3K36me2, known to inhibit transcriptional initiation (Strahl et al. 2002; Li et al. 2009), is found to be depleted in CGI chromatin compared to the non-CGIs promoters and gene body. Depletion of H3K36me2 was also found to be linked to the appearance of the H3K36me3 state (Blackledge et al. 2010) which is a gene body histone mark and associated with actively transcribed genes (Mikkelsen et al. 2007). The histone demethylase Kdm2a, a CxxC domain protein that binds particularly to unmethylated CpG, mediates demethylation of H3K36me2 (Tsukada et al. 2006; Blackledge et al. 2010). Hence, reduction of H3K36me2 may contribute to a transcriptionally permissive state at CGIs. Both modifier proteins Cfp1 and Kdm2a depend on CpG density to influence chromatin modification (Figure 1.3).

Repressive state

CGIs promoter can be silenced by DNA methylation and polycomb-mediated mechanism (Figure1.3). CGIs are not methylated when located at transcription start sites (TSSs). Usually CGIs are unmethylated compared to the other heavily methylated genome. It is well known that some CGIs

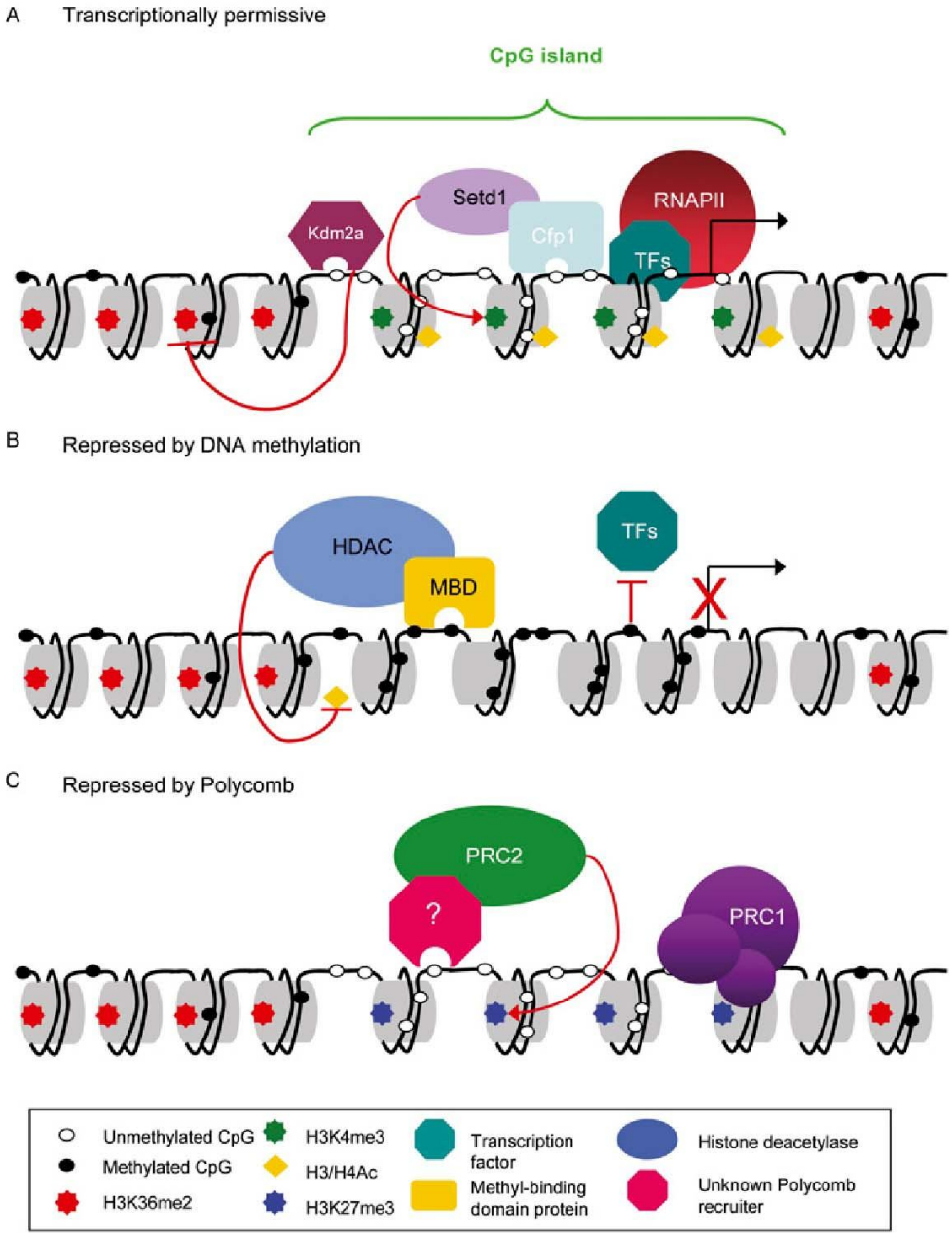

Figure 1.3. The chromatin state at CGIs. (A) Transcriptionally permissive unmethylated CGIs, marked by histone acetylation (H3/H4Ac) and H3K4me3, which is directed by Cfp1, and show Kdm2a-dependent H3K36me2 depletion. Nucleosome deficiency and constitutive binding of RNAPII may also contribute to this transcriptionally permissive state. (B) DNA methylation associated stable long-term silencing of CGI promoters mediated by MBD proteins, which recruit corepressor complexes associated with HDAC activity, or may be due to directed inhibition of transcription factor binding by DNA methylation. (C) Polycomb mediated silencing of CGIs. An unknown CGI-binding factor could be responsible to recruit PRC2 to CGIs that trimethylates H3K27. This H3K27me3 is recognized by PRC1 complexes that act to inhibit transcriptional elongation, thereby silencing genes. (Deaton and Bird 2011).

acquire methylation during normal differentiation, which leads to the stable silencing of the associated promoter (Stein et al. 1982; Mohn et al. 2008; Payer and Lee 2008). DNA methylation can destroy permissive chromatin environment by inhibiting the ZF-CxxC domain proteins that specifically recognizes CpG dinucleotides (Blackledge et al. 2010; Thomson et al. 2010). Also methylated CpG acts as binding sites for methyl CpG binding domain proteins (MBDs) that recruit repressors of transcription, histone deacetylase (HDAC) complexes (Figure 1.3) (Meehan et al. 1989; Nan et al. 1998; Klose and Bird 2006; Clouaire and Stancheva 2008).

The mechanism of susceptibility of this CGI subset to DNA methylation is not still clear. Studies from cancer tissues reported that a polycomb-silenced intermediate state may facilitate the constitutively repressed and acquisition of DNA methylation state (Viré et al. 2005; Schlesinger et al. 2006). Unlike CGIs at the promoter of annotated genes, orphan CGIs are frequently methylated. CGIs in gene bodies are sometimes methylated in a tissue-specific manner (Illingworth et al. 2010; Maunakea et al. 2010; Deaton et al. 2011). Elevated methylation of intragenic CGI is found to be correlated with silencing of the associated gene.

In addition to DNA methylation, polycomb group proteins (PcG) mediate silencing of CGI promoters through two distinct complexes: polycomb repressive complex 1 (PRC1) and PRC2. PRC2 mediates H3K27me3, and this mark is recognized by PRC1, which is thought to inhibit transcriptional elongation, thus contribute to genes silencing (Figure 1.3). It is noteworthy that the transcriptionally permissive and polycomb-repressed states can coexist at CGIs with bivalent domain, predominantly in totipotent embryonic cells (Deaton and Bird 2011).

Epigenetics and evolution

The role of epigenetic modifications in the regulation of gene expression is crucial in eukaryotic biology as well as in evolution. Epigenetic states can be influenced by environmental cues to affect phenotype for multiple generations. Epigenetic changes in gene expression not only creates heritable phenotypic diversity within an individual, but also within populations, independent of genetic variation (Richards 2006; Richards 2008). Changes in the regulation of gene expression levels have long been thought to play a vital role in evolution and adaptation (Britten and Davidson 1969). It has been hypothesized that epigenetic variant could be a novel substrate for natural selection and thus participate in the adaptation of species in changing environment (Jablonka and Lamb 1989).

Several studies have reported the association between DNA methylation and epigenomic variation. In higher eukaryotes, individuals largely differ each other in their epigenomic chromatin signatures. Intra-species epigenomic variation has been found at DNA methylation level in *Arabidopsis thaliana* (Vaughn et al. 2007; Zhang et al. 2008). A recent study showed progressively increased prevalence of an epigenetic trait in mice (dietary methyl donors supplemented) in the population over five generations. Withdrawal of the dietary supplement resulted loss of that trait after one generation, supports the idea that beside genetic variation natural selection can act on epigenetic variation. Their finding also suggests that epigenetic changes could underlie rapid adaptation of species in response to natural environmental change (Cropley et al. 2012). In humans, several studies reported the inter-individual differences of DNA methylation (Fraga et al. 2005; Flanagan et al. 2006; Gibbs et al. 2010; Zhang et al. 2010). Another study showed DNA methylation patterns are associated with genetic and gene expression variation in human cell lines (Bell et al. 2011). Comparative studies found significant interspecies methylation level differences across tissues between human and other primates (Gama-

Sosa et al. 1983; Enard et al. 2004). Llamas et al. showed the stability of cytosine methylation patterns in DNA from ancient specimens by bisulphite allelic sequencing of loci from late Pleistocene bison and suggested that cytosine methylation in ancient DNA provides a powerful means to study the role of epigenetics in evolution (Llamas et al. 2012). Whole-genome sequencing of other ancient samples has also been demonstrated (Miller et al. 2008; Green et al. 2010; Reich et al. 2010) that open the opportunity to explore precisely the correlation of DNA methylation patterns with evolution over evolutionary time-frame.

Beside DNA methylation, histone modifications could also contribute to the epigenetic variation. In a comparative study, 15 genomic regions in humans associated with histone acetylation were found to be conserved for the same epigenetic status in 10 of the orthologous regions in mouse (Roh et al. 2007). On a larger scale comparison for different histone modifications between human and mouse for the chromosomes 21 and 22 and the syntenic chromosomes, reported that genomic locations of those epigenetic markers at orthologous loci are strongly conserved, even in the absence of sequence conservation. Interestingly, the conservation of histone modification patterns was highest in genomic regions proximal to annotated orthologous genes (Bernstein et al. 2005; Wilson et al. 2008). Cain et al. proposed that differences in gene expression levels among primates are associated with the changes in H3K4me3 (Cain et al. 2011). A recent study identified human-specific changes in H3K4me3 levels at TSSs and related regulatory sequences in comparison with chimpanzees and macaques (Shulha et al. 2012). Using a chemical perturbation of yeast cell model, a recent study found the persistence of acetylation variation at some nucleosomes, stabilized by a DNA variance and alteration in other nucleosomes, might have experienced environmental perturbations (Abraham et al. 2012).

Genetic information also affects gain and loss of epigenetic marks in the form of either cis- or transacting variation. For instance, trans-acting genetic variation in the genes that encode enzymes (e.g. DNA methyltransferase or histone methyltransferase) responsible to add epigenetic marks, could contribute to epigenetic variation, while cis-acting variation in the nucleotide sequence at the target locus might also play a role in epigenetic regulation (Murrell et al. 2004; Heijmans et al. 2007).

Rapid production of genomic and epigenomic data will facilitate refine our understanding of epigenetic mechanisms and connections between these mechanisms and evolutionary change at the population level.

Selective pressure and signatures of natural selection

Selective pressure is the phenomena which alters the behaviour and fitness of living organisms within a certain environment. It is the driving force of natural selection and evolution. The recent history of the human population is characterized by huge environmental change and emergent selective agents (Sabeti et al. 2002). Dramatic changes in environment and lifestyle likely resulted in powerful selective pressures for new genotypes that were better suited to the novel environments (Pickrell et al. 2009).

The ability to identify the molecular signature of natural selection provides a powerful tool for identifying loci that have contributed to adaptation. Several methods have been developed that identify signals from both recent and ancient selection event in human genome on a genome-wide scale (Voight et al. 2006; Sabeti et al. 2007; Pickrell et al. 2009; Green et al. 2010; Pollard et al. 2010). These tools could further facilitate the study of the relationship between epigenetic modification and

selective pressure in a genomic perspective. Since one of the main effects of selection is to modify the levels of variability within and between species, these methods could be roughly classified into two groups: the methods in the first group use a population genetic approach, while the second group use a comparative approach. Population genetic approaches are mainly used to detect recent selection events occurring in a population, comparative approaches, on the other hand, deal data from multiple different species, are suitable to detect more ancient selections (Nielsen 2005).

Aims

Considering the crucial role of epigenetics modifications in gene regulation, our aim was to explore the relationship between epigenetic modifications and selective pressure in human genome. In particular:

1. In the first step we focused the relationship between cytosine methylation in CpG islands and signatures of selective pressure in human genome.

2. In the second step we focused to explore the relationship between histone modifications (HeK3me3, H3K27ac and H4K36me3) and signatures of selective pressure in human genome.

Chapter 2: CpG islands are undermethylated in the genomic regions under selective pressure

Introduction

DNA methylation at CpG sites is one of the most intensively studied epigenetic mechanisms (Pelizzola and Ecker 2011). CpG sites are DNA regions where a cytosine nucleotide occurs next to a guanine nucleotide. Cytosines in CpG dinucleotides can be methylated to form 5-methylcytosine. Human genome contains about 30 million CpGs that exist in a methylated or unmethylated state. A part of all CpG sites present in the genome are clustered into CpG islands that are defined as genomic regions with increased CpG density. These CGIs are enriched at genes, about 60% of all genes in the human genome containing a CpG island upstream (Bird 2002). The methylation status of CGIs can influence gene expression (Illingworth and Bird 2009; Pelizzola and Ecker 2011). The hypermethylation at promoter CGIs typically results in a decreased transcription of downstream genes (Stein et al. 1982). Further, aberrant DNA methylation has been often reported to cause various human diseases (Handel et al. 2010; Petronis 2010; Duthie 2011).

Three DNA methyltransferases, namely DNMT1, DNMT3a, and DNMT3b (Jones and Liang 2009) are involved in the maintenance of DNA methylation during the cell cycle. When the two parental DNA strands are separated in the S-phase of the mitosis, two hemimethylated strands are produced. DNMT1 is a component of a protein complex with high affinity with hemimethylated DNA, subsequently restoring methylation on the daughter strands (Sharif et al. 2007).

Also demethylation is an important biological mechanism, as illustrated, for example, by the demethylation of the paternal and maternal genomes in the zygote after fertilization (Haaf 2006) or by the reprogramming of pluripotency cells to differentiated cells (Mikkelsen et al. 2008). Nevertheless, the molecular mechanism of DNA demethylation in mammals is disputed, one possibility being that cells demethylate their genome by passive demethylation.

Several evidences suggest a dependence of DNA methylation on local sequence content (Bock et al. 2006). DNA methyltransferases within eukaryotic cells are not free, but they are compartmentalized by interaction with nuclear components (Jeong et al. 2009). Thus it is likely that chromatin structure of a genomic region will have an important impact on the maintenance of methylation of that region. It could be hypothesized that there are genomic regions somehow ‘‘protected’’ in vivo from methylation but yet readily accessible to exogenously added soluble DNA methylases (Lin et al. 2007). Nonetheless, a complete understanding of the role of DNA methylation and the mechanisms responsible for its establishment and maintenance remain elusive (Pelizzola and Ecker 2011).

Many studies focused on the interplay between epigenomic regulation and evolution, because DNA methylation is involved in several processes that are central to evolutionary biology, including phenotypic plasticity and evolvability (Johnson and Tricker 2010). Changes in the regulation of gene expression levels have long been hypothesized to play an important role in evolution (Britten and Davidson 1969). Nevertheless, studies specifically addressed to the relation between promoter methylation and selective pressure in *Homo sapiens* are still lacking.

Several tools are needed to study the relation between CGIs methylation and selective pressure in a genomic perspective. First, we need tools that recognize genomic signals of selective pressure. Many methods have been developed to exploit signatures left by natural selection, each signature providing distinct information about selective events (Nielsen 2005). Since one of the main effects of selection is to modify the levels of variability within and between species, these methods could be roughly classified into two groups. To the first group belong the methods that use a population genetic approach, while to the second group belong methods that use a comparative approach. While population genetic approaches aim to detect recent selection events occurring in a population, comparative approaches, involving data from multiple different species, are suitable for detecting more ancient selections (Nielsen 2005). By these methods, hundreds of such regions putatively under selective pressure have been identified. They are typically as large as few hundreds of kilobases to megabases, and may contain many genes. The second requirement to study the relation between CGIs methylation and evolution is the availability of methylation data at genomic scale. Recent advances in high-throughput sequencing technologies are enabling epigenetics to progress rapidly into an 'omic' science (Fouse et al. 2010). In particular, the Encyclopedia of DNA Elements (ENCODE) Consortium (Meissner et al. 2008; Celniker et al. 2009) is providing masses of methylation data that may be accessed and used by the entire scientific community. The analysis of these relevant datasets by computational methods could complement experimental approaches to further our understanding of DNA methylation (Bock and Lengauer 2008; Soojin and Goodisman 2009).

In this study, we explored the relationship between CGIs methylation and signatures of selective pressure in *Homo sapiens*, using a computational methodology. We compared the CGIs methylation level in genomic regions under selective pressure with CGIs localized in the remaining genome. We evaluated DNA methylation levels both by direct analysis of CpG methylation in cell lines and by an indirect approach that uses the analysis of genetic variation inside CGIs. To define genomic regions under selective pressure, we used three different methods oriented to provide information about selective events happened in different periods of human evolution. Independently of the methods used both to evaluate CGIs methylation and to estimate selective pressure, we found evidences of undermethylation of CGIs in human genomic regions that undergone selection.

Results

DNA methylation in cell lines and signatures of selective pressure

Based on datasets available in public repository we estimated the CGIs methylation in 25 cell lines. Genomic coordinates of 28691 CGIs were obtained from UCSC Genome Browser ''CpG Islands'' track. As known, USCS CGIs file contains also data related to sequence for alternative haplotypes (present mainly in chr 6, for the inclusion of alternative versions of the MHC region). Of course, in our analysis we filtered the file excluding these duplicated data. Excluding CGIs corresponding to sequences for alternative haplotypes, we obtained 27718 unique CGIs.

Cell line methylation data were obtained by downloading them from UCSC Genome Browser ''HAIB Methyl RRBS'' track. This track reports the percentage of DNA molecules that show cytosine methylation at specific CpG dinucleotides in several cell lines. The 25 cell lines that we used could be roughly divided in three groups: cancer transformed cells (n= 6), EBV transformed cells (n= 2) and normal untransformed cells (n= 17). The complete list of the cell used, with their characteristics are shown in Supplimentary table 2.1. We extracted only the methylation values of those CpGs that were localized inside CGIs (order 10^5 per cell line).

To estimate the methylation of each CpG island we calculated the mean of all CpGs methylation values into a CpG island. We were able to estimate the methylation status of about 10^4 CGIs for each cell line. Supplimentary table 2.2 lists, for each cell type, the description of the CpGs analyzed. As expected, the CGIs mean methylation values were higher in Cancer Transformed (mean = 26.91, SE = 2.84) and lower in Normal Untransformed cells (mean = 14.34, SE = 0.57), EBV transformed cell showing intermediate levels (mean = 18.93, SE = 1.46) (Figure 2.1).

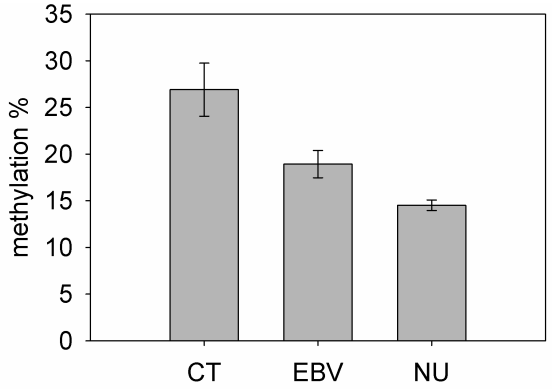

Figure 2.1: Histogram of CGIs mean methylation values (y-axis) and their Standard Errors for each cell line group: Cancer Transformed (CT), EBV transformed (EBV), and Normal Untransformed (NU).

To explore the possible relationship between CGIs methylation and selective pressure we compared the methylation of the CGIs inside genomic regions showing signature of selective pressure with the methylation of the CGIs in the remaining genomic regions.

To obtain genomic regions with signatures of selective pressure, we used three different approaches. As first approach, we used the per-continent Integrated Haplotype Score (iHS) (Voight et al. 2006)**.** This score belongs to the Extended Haplotype Homozygosity (EHH) statistic ''family'' (Sabeti et al. 2002). The iHS measures the decay of identity, as a function of distance, of haplotypes that carry a specified ''core'' allele at one end and it is considered a measure of recent positive selection. The normalized iHS scores (see materials and methods) were obtained from UCSC Genome Browser ''HGDP iHS'' track.

To define genomic regions putatively under selective pressure by this method, we scanned normalized iHS scores across the whole genome and selected the genomic intervals where iHS score values ≥ 2. Once detected such compact regions, we extended their boundaries to the nearest loci where iHS was exactly vanishing. According to these criteria, 586 regions were identifies. We denoted these regions as ''High iHS regions'' (HIR). Supplimentary table 2.3 reports the HIRs that we identified and their boundaries. Next we identified CGIs localized within HIRs. We found that 2545 CGIs were localized inside HIRs whereas the remaining 26146 were placed outside. We compared the methylation of CGIs inside HIRs with the methylation of CGIs localized outside these regions.

Figure 2.2 shows the results obtained. In all cell lines analyzed, the CGIs inside HIR regions were less methylated than the CGIs in the remaining part of the genome. The differences were highlystatistical significant (Bootstrap p-value$\leq 10^{-4}$) in all cell lines analyzed. Supplimentary table 2.4 reports in detail

the results of this analysis. The Bootstrap procedure adopted to evaluate the difference between means of distributions is described in Materials and Methods.

An additional method able to detect regions putatively under selective pressure is represented by the Selective Sweep Scan (S) score, which is based on the comparison of Homo Sapiens DNA with Neanderthal DNA (Green et al. 2010). This score, when positive, indicates more derived alleles in Neanderthal than expected, given the frequency of derived alleles in human. On contrary, a negative score indicates fewer derived alleles in Neanderthal, and may suggest an episode of positive selection in early humans, after divergence with Neanderthal and before human populations divergence. We used the 212 regions with S scores in the lowest 5% of the distribution (5% Lowest S Regions, 5LSR) contained in the UCSC Genome Browser (see materials and methods). Supplimentary table 2.5 reports the regions used with their relative boundaries.

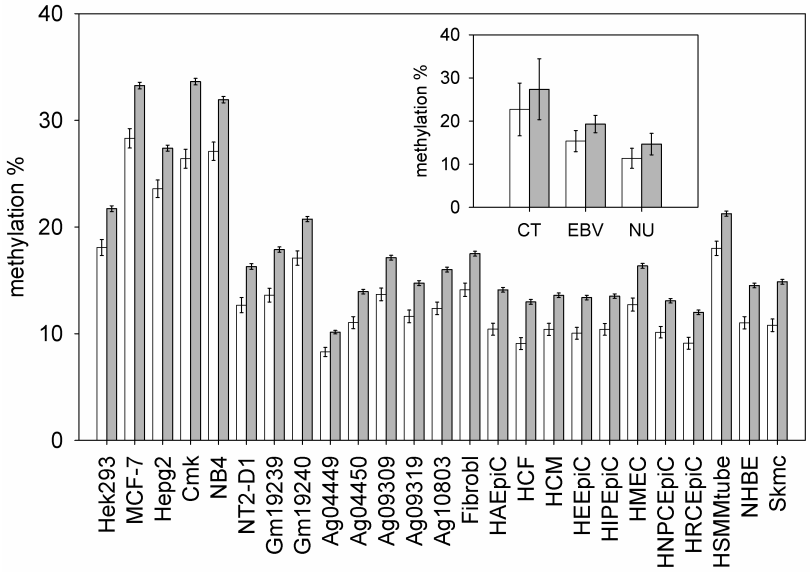

Figure 2.2: **Methylation of HIR CGIs compared to methylation of CGIs in other genomic regions.** For each cell line, the mean methylation value of CGIs inside HIR regions (open bars) and of the CGIs in the remaining part of the genome (closed bars) are reported. Inset shows the same data summarized by cell group (Cancer Transformed = CT, EBV transformed = EBV, Normal Untransformed = NU). Values are means +/− Standard Error (SE).

We found that 348 CGIs were localized inside 5LSRs and the remaining 28343 outside them. Figure 2.3 shows the results obtained by comparing the methylation of CGIs inside 5LSRs with the methylation of CGIs localized in the other regions of the genome.

Also for this different measure of selective pressure, in all cell lines analyzed, CGIs inside regions under selective pressure were less methylated than the remaining CGIs. The differences were highly statistical significant (Bootstrap p-value<10^{-3}) in 17 cell lines analyzed, but did not reach this significance in 8 cell lines ($p<0.05$). Nevertheless, combining the results of all 25 cell lines by means of the test statistic - 2 log (p1, p2 ... p25), where p1, p2 ... p25 are the p-values of the individual tests, we reached a combined statistical significance much less than 10^{-3}. Supplimentary table 2.6 reports in detail the results of the analysis.

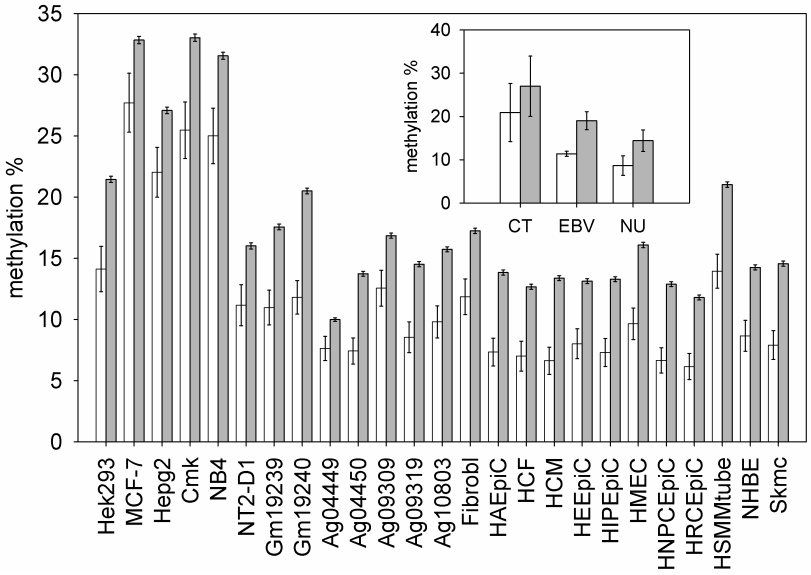

Figure 2.3: Methylation of 5LSRs CGIs compared to methylation of CGIs in other genomic regions. For each cell line, the mean methylation value of CGIs inside 5LSRs regions (open bars) and of the CGIs in the remaining part of the genome (closed bars) are reported. Inset shows the same data summarized by cell group (Cancer Transformed = CT, EBV transformed = EBV, Normal Untransformed = NU). Values are means +/− SE.

To check if the results could be due to the same CGIs identified by both methods, we searched for CGIs that are both within HIRs and within 5LSRs. We found only 70 CGIs in common between these two groups, indicating that the results obtained by the two methods are driven by different sets of CGIs. In addition, excluding these 70 CGIs from the analysis, the result continued to be highly significant both for HIRs and 5LSRs (data not shown). It is intriguing to note that these 70 CGIs were less methylated when compared both to the remaining HIR CGIs and 5SLR CGIs, but the differences were not statistical significant (data not shown).

To further define regions under selective pressure, we decided to use a third and last approach that looks for sequences that are conserved across species (Pollard et al. 2010). By this approach, conserved regions are defined as genomic regions with a reduced rate of evolution compared to what is expected under neutral drift. Several methods for detecting conserved regions in multiple alignments have been described. We used data downloaded from UCSC Genome Browser Conservation (cons46way) Track, which lists 725627 Conserved Elements (CEs) that were predicted to be conserved among primates (Siepel et al. 2005).

We Identified 26936 CEs located inside 14391 CGIs, by filtering all genomic CEs by CGIs. Excluding CGIs corresponding to sequences for alternative haplotypes, we obtained 13288 unique CGIs containing 25362 CEs. We named ''CE CpG islands (CE CGIs)'' those CGIs that contain at least one conserved element. For each cell line, we compared the methylation of CE CGIs with the methylation level of the remaining CGIs not containing conserved elements.
In all the cell lines analyzed, CE CGIs were less methylated than CGIs that do not contain conserved elements (Figure 2.4). The differences were highly statistical significant (Bootstrap p-value<10^{-4}) in all lines analyzed. Supplimentary table 2.7 reports in detail the results of this analysis.

Since the number of CE CGIs is higher than that of HIR CGIS and 5SLR CGIs, it could be possible that all HIR CGIs and 5SLR CGIs are contained in the CE CGI group. In this case the results we found with HIR and 5SLR could be due to CE only.

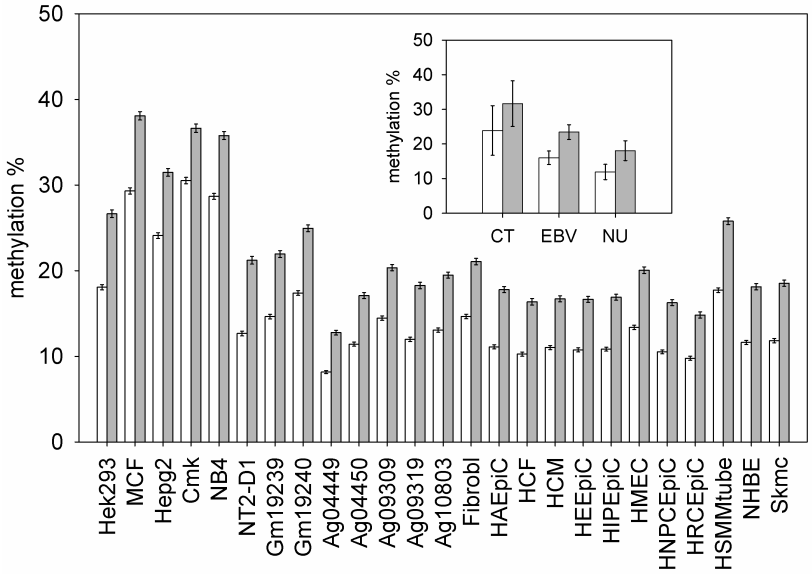

Figure 2.4: Methylation of CE CGIs compared to methylation of CGIs that do not contain conserved elements. For each cell line, the mean methylation value of CE CGIs (open bars) and of the CGIs that do not contain conserved elements (closed bars) are reported. Inset shows the same data summarized by cell group (Cancer Transformed = CT, EBV transformed = EBV, Normal Untransformed = NU). Values are means +/– SE.

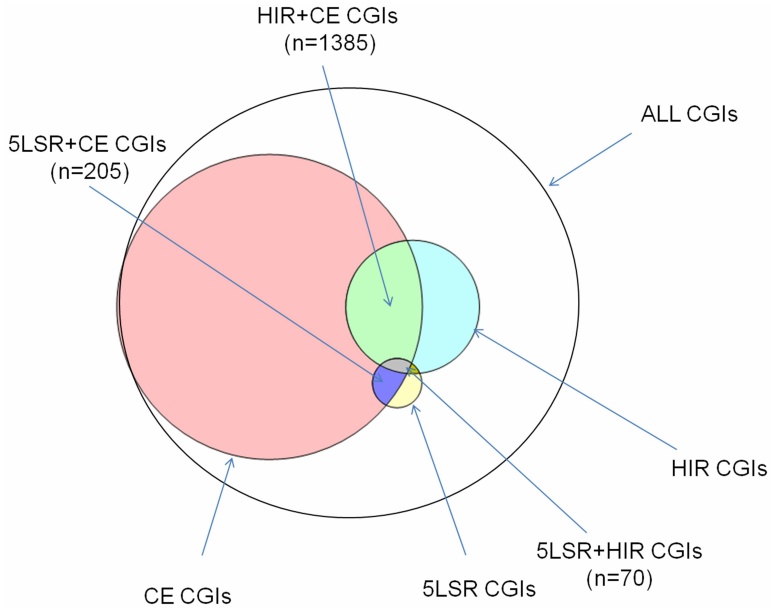

Figure 2.5: Venn diagram showing the overlaps among CGIs localized in the regions under selective pressure detected by the three methods used.

To check this possibility, we estimated the overlaps between the CGIs lists obtained by the different methods (Figure 2.5). We found that 1385 CGIs were in common between CE and HIR (HIR+CE CGIs) and 205 were in common between CE and 5SLR (5SLR+CE CGIs).

If the phenomena underlying the three signatures (CE, HIR and 5SLR) contributed independently to lower the CGIs methylation, we expected that CGIs in regions with two signatures of selective pressure showed lower methylation when compared to CGIs in regions with one signature only.

We found that, in all cell lines analyzed, HIR+CE CGIs were less methylated than the remaining CE CGIs. The differences were highly statistical significant (Bootstrap p-value< 10^{-3}) in 14 cell lines analyzed, but did not reach this significance in 11. In these eleven cell lines the differences were significant only at p<0.05 (Figure 2.6, Supplementary table 2.8).

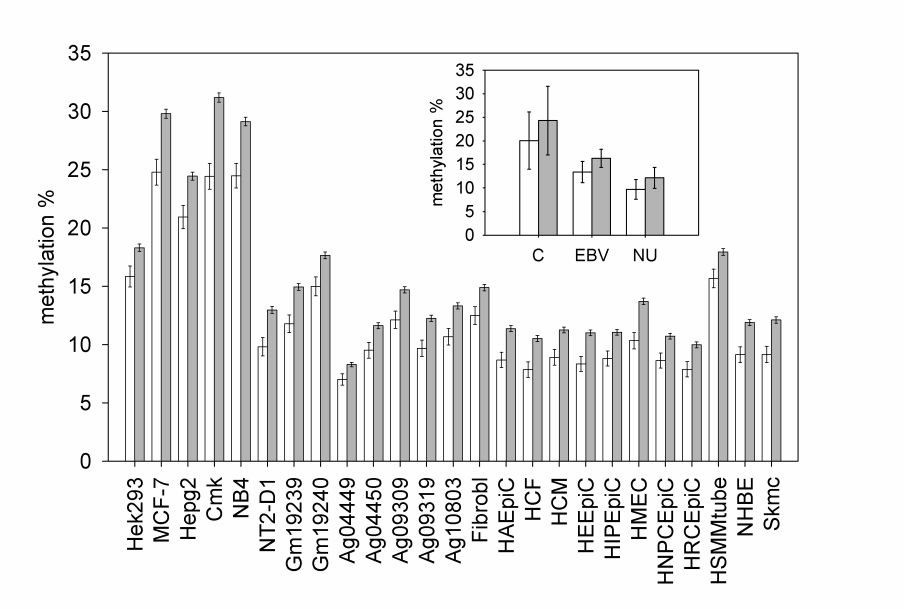

Figure 2.6: Histogram of the percentages of methylation of HIR+CE CGIs (open bars) compared to CE CGIs (closed bars) for each cell line. Error bars represent standard errors.

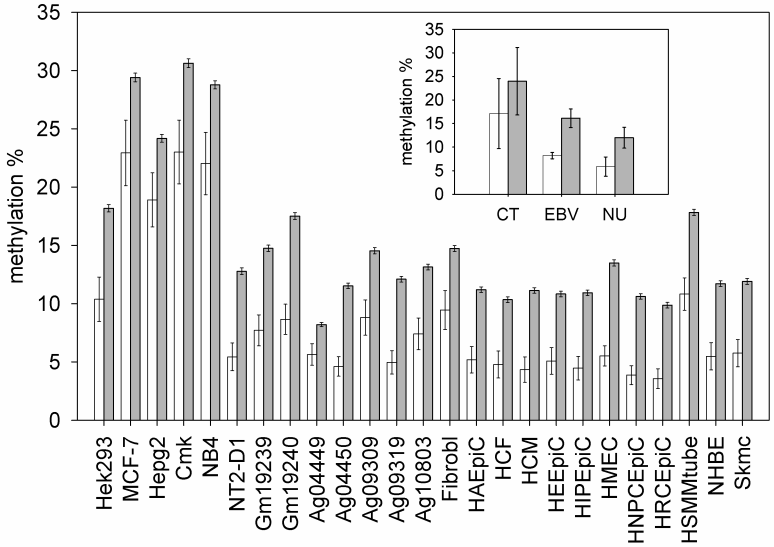

Figure 2.7: Histogram of the percentages of methylation of 5SLR+CE CGIs (open bars) compared to CE CGIs (closed bars) for each cell line. Error bars represent standard errors.

Also 5SLR+CE CGIs were less methylated when compared to the remaining CE CGIs, in all cell lines analyzed. The differences were highly statistical significant (Bootstrap p-value, 10^{-3}) in 17 cell lines, but did not reach this significance in 8. In these eight cell lines the differences were significant only at $p<0.05$ (Figure 2.7, Supplementary table 2.9). Also in these two cases the joint analysis of all cell lines yielded a combined statistical significance much less than 10^{-3}.

In the genome, CGIs are located in 5', 3' or in other gene regions, as well as in intergenic regions. We decided to estimate the methylation of CGIs located in these different locations to assess if the CGIs undermethylation that we found in regions under selective pressure is restricted to CGIs with a specific localization. We used the 4 classes of CpG islands described by Medvedeva et al. (Medvedeva et al. 2010): 5' CGIs (in 5'-flank region, 5' UTR-exon , 5'UTRintron, initial coding exon and initial intron), intragenic CGIs (in internal exons and internal introns), 3' CGIs (in final exons, final introns, 3' UTR exons and 3' UTR introns) and intergenic CGIs (located at least 3 kb from any known gene upstream and downstream). In particular, 5' CGIs are located in regions that, starting 3 kb upstream Transcription Start Site, extend till the first intron. Considering all cell lines, 5' CGIs showed the lowest methylation level (weighted mean = 9.01), intragenic and 3' CGIs showed the highest values (respectively, weighted mean = 55.21 and 42.59) and intergenic CGIs showed intermediate methylation values (weighted mean = 21.31). For each cell line, the differences among CGIs methylation of different genomic regions were high statistical significant (Kruskal-Wallis Test, p-value $\leq 2.2\ 10^{-16}$) (Supplementary table 2.10).

Next we divided CGIs with signature of selective pressure according the above described classes. Unfortunately, for intragenic and 3' classes, we did not obtain a number of HIR CGIs and 5LSR CGIs sufficient to perform a consistent statistical analysis. In particular, in these classes we found about 80 HIR CGIs and less than 10 5LSR CGIs.

We were able to perform statistical analysis only by using CE as signature of selective pressure. In all cell lines but 2 (which were both cancer cell lines), 5' CGIs in CE regions were undermethylated when compared to 5' CGIs located outside CE regions (Bootstrap p-value,10^{-4}). Intragenic and 3' CGIs located in CE regions showed no differences in methylation when compared to intragenic and 3' CGIs outside CE regions. In all cell lines, intergenic CGIs in CE regions were severely undermethylated when compared to intergenic CGIs located outside CE regions (Bootstrap p-value $<10^{-4}$) (Supplementary table 2.11).

This first set of experiments suggested that, in different cell lines, the GCIs localized in genomic regions under selective pressure are undermethylated. CGIs in regions with two signatures of selective pressure (in which CE is involved) showed lower methylation when compared to CGIs in regions with one signature only. Furthermore, at least for CE, the CGIs undermethylation that we found in genomic regions under selective pressure is specifically due by CGIs located at the 5' and in the intergenic regions.

Genetic variation inside CpG islands and signatures of selective pressure

We decided to estimate the CGIs methylation by a different, indirect approach. It is well settled that 5-methylcytosine is the initial molecule in the deamination reaction that generates thymine; thus, methylation may be required for increased mutation rates at CpG sequences. We predicted that CGIs

localized in regions under selective pressure, being less methylated, would be less likely to mutate. Under this hypothesis, these CGIs should show a lower degree of genetic variation among individuals.

To evaluate the degree of genetic variation in CGIs, we calculated the frequency of SNPs present in each CGI. Among the 26033053 SNPs from dbSNP (build 131), we selected the 199514 SNPs that were located inside CGIs. To obtain a normalized value of SNP frequency for each CGI, we divided the number of SNPs present in each CGI by its size. By this method we were able to calculate the SNP frequencies for 25558 CGIs.

We found that, on average, each CGI contained 1.04 SNP/100 bp (range 0.04–63.28). Then we compared the SNP frequency of CGIs inside the regions under selective pressure with the SNP frequency of CGIs localized in the other regions of the genome.

Figure 2.8 reports the results obtained. The 2345 CGIs localized in HIRs showed a mean of 0.89 SNP/100 bp in comparison with 1.05 of the other 23213 CGIs (Bootstrap p-value < 10^{-4}).

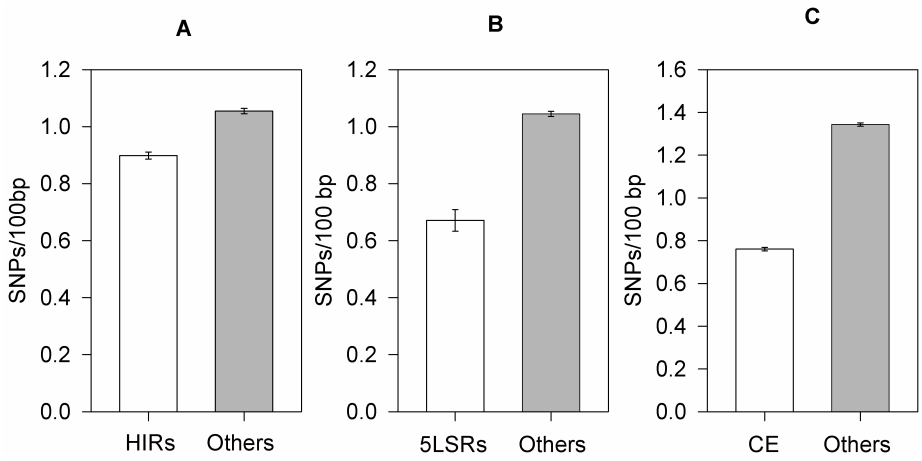

Figure 2.8: SNP content of CGIs in genomic regions under selective pressure compared with CGIs localized in other genomic regions. The mean SNP frequencies (SNPs/100 bp) of CGIs in genomic regions under selective pressure (open bars) and of CGIs localized in other genomic regions (closed bars) are reported. The regions are: A = HIR, B = 5SLR and C = CE). Values are means +/– SE.

The 309 CGIs localized in 5LSRs showed a mean of 0.67 SNP/100 bp in comparison with 1.04 of the other 25249 CGIs (Bootstrap p-value < 10^{-4}). The 13286 CE CGIs showed a mean of 0.76 SNP/100 bp in comparison with 1.34 of the other 12272 CGIs (Bootstrap p-value < 10^{-4}).

Also for this approach we checked if CGIs in regions with two signatures of selective pressure (HIR+CE or 5SLS+CE) showed differences compared to CGIs in regions showing only a signature (CE). We found that the 205 5SLR+CE CGIs contained less SNPs than the remaining 13081 CE CGIs (mean= 0.61 SNP/100 bp vs. 0.76 SNP/100 bp, Bootstrap p-value < 10^{-4}). On contrary, the 1386 HIR-CE CGIs did not show any difference in SNP content in comparison with the remaining 11900 CE CGIs (mean= 0.74 vs. 0.76, Bootstrap p-value = 0.36).

In summary, we demonstrated that the CGIs localized in regions showing signatures of selective pressures contain less SNPs than CGIs in other regions of the genome. When compared to CGIs in

regions with one signature only, CGIs in regions with two signatures of selective pressure showed differences in the case of 5SLR but not for HIR.

Discussion

One of the most studied epigenetic modifications is the DNA methylation, which is evolutionarily ancient and associated with regulation of gene transcription (Law and Jacobsen 2010). DNA methylation could be central both to the ability of a population of organisms to change its phenotype in response to changes in the environment and to its ability to generate genetic diversity and evolve through natural selection (Johnson and Tricker 2010). The evolutionary conservation and divergence of epigenetic mechanisms in eukaryotes have started to be revealed by genetic and genomic studies of various organisms (Law and Jacobsen 2010). A general scenario that seems to emerge is that the epigenetic marks and the mechanisms that establish these marks are basically ancient and conserved, but the precise details of how these marks function within genomes is far to be completely clarified. An intriguing question is how evolutionary forces have adapted epigenetic mechanisms to the needs of the specific organism and, within a species, to the needs of a specific population.

In this study we searched for possible differences in DNA methylation between genomic regions under selective pressure and the remaining genome. We focused on CpGs inside CpG islands and on the species *Homo sapiens*. We chose a genome-wide approach using computational biology methods.

One of the difficulties in this kind of study concerns the methods to be used to detect signatures left by natural selection. Despite the many methods that have been developed, up to now no method could be considered the ''best one''. Each method apparently provides distinct information about selective events (Nielsen 2005). To overcome this limit we decided to use three different approaches. The first one, the iHS score (Voight et al. 2006), is a population genetic approach. The general idea of this method is to search for haplotypes longer than expected, the so-called ''long-range haplotype''. An allele under selection increases in frequency so rapidly that long-range associations with neighboring polymorphisms are not disrupted by recombination. Generally this approach is thought to provide evidence for recent positive pressure (Voight et al. 2006), ''recent'' meaning after the human population separation. The second method defines as ''under selective pressure'' the regions of the human genome with a strong signal for depletion of Neanderthal-derived alleles. The presence of these signals may mark an episode of positive selection in early humans, after the separation from Neanderthal (Green et al. 2010). The third and last method belongs to the comparative approaches, involving data from multiple different species. Methods for detecting signatures of selection from rates and patterns of substitution have a long history in the field of molecular evolution (Petronis 2010). The method that we used (Pollard et al. 2010) is aimed to identify conserved elements in primates allowing to test hypotheses about selective pressures on this particular evolutionary lineage. We decided to use these three methods because they provide information about selective events happened in different evolutionary times.

Independently of the method that we used, CGIs localized inside regions under selective pressure were less methylated than CGIs in other genomic regions. In addition, we found that CGIs in regions with two signatures of selective pressure (in which CE is involved) showed lower methylation when compared to CGIs in regions with one signature only. This finding suggests that each signature is providing distinct information about selective events.

We observed CGIs undermethylation in all cell lines analyzed, including different types of normal cultured cells (fibroblasts, epithelial cells, myocytes etc.). It is well known that, in a multicellular organism, different cell types acquire various functional capabilities by distinct epigenetic modifications. Acquired during early development, the cell type-specific epigenotype is maintained by cellular memory mechanisms. It is quite surprising that different cells showed similar methylation differences. This finding may suggest that the regions under selective pressure are somehow more ''protected'' from methylation, independently of the cell type-specific epigenotype. This interpretation could be further supported by the analysis of EBV transformed and cancer derived cells. Epigenetics of cancer has been deeply studied, and the loss of DNA methylation at CpG dinucleotides was the first epigenetic abnormality to be identified in cancer cells (Feinberg and Tycko 2004). The role of hypomethylation in activating oncogenes, as well as hypermethylation affects tumor-suppressor genes has been well established (Feinberg and Tycko 2004). We found that genomic regions under selective pressure are relatively less methylated in cancer cells too. This difference persists even in a scenario of global hypermethylation that characterizes cancer cells in our experiments.

To confirm the results obtained in cell lines, we checked the possible existence of undermethylation in regions under selective pressure by a different approach. It is well established in scientific literature that the 5methylcytosine present in some CpG sites is subject to mutational pressure by spontaneous deamination to thymine (Holliday and Grigg 1993). A fraction of CpG sites in the genome are clustered into CpG islands that are thought to be mainly unmethylated (Bird 1986). Since 5-Methylcytosine is the initial molecule in the deamination reaction that generates thymine, CpG sequences within CpG islands, which are not methylated, would be less likely to mutate. Tomso et al. (Tomso and Bell 2003) found a general underrepresentation of polymorphisms in CpG islands, strongly supporting the idea that decreased methylation in CpG islands leads to decreased variation at island CpGs. Using the same way of reasoning, we predicted that, if CGIs in regions under selective pressure were undermethylated, they would show less polymorphisms than the CGIs in the remaining genome.

Independently of the method used to define the regions under selective pressure, we found that CGIs inside regions under selective pressure contain less SNPs than the CGIs in the remaining genome. When we compared CGIs in regions with two signatures of selective pressure to CGIs in regions with one signature only, we found that CGIs showing both 5SLR and CE signatures contained less SNPs than CGIs showing CE signature only. On the contrary, when we compared CGIs showing both HIR and CE signatures to CGIs showing CE signature only, we found no differences in SNP content. A possible explanation is that the selective pressure that acted on HIRs was very recent. Its effect could be evident in cell CGIs methylation but not (or not yet) in genetic variation.

CGIs can be located inside the genes or outside them. CGIs located inside genes can be divided, according their position, in CGIs in 5' region, CGIs in the 3' regions and CGIs in internal exons or introns. CGIs located near 5' region of genes are known to influence gene expression but also CGIs located outside these regions can be involved in important biological processes (Smilinich et al. 1999; Ramser et al. 2008; Illingworth and Bird 2009). We decided to analyze the methylation of CGIs, categorized by their position, to assess if the CGIs undermethylation that we have found in regions under selective pressure was a general phenomenon or it was restricted to CGIs with a specific localization. We were able to analyze only CE CGIs because, after classification, the number of HIR CGIs and 5LSR CGIs in intragenic and 3' regions was too low to perform a reliable statistical analysis.

We found that, at least for CE, the CGIs undermethylation in regions under selective pressure specifically involved CGIs located at the 5' and in the intergenic regions. For the 5' regions, the finding was quite expected because of their well established role in gene regulation. The functional role of intergenic CGIs is less clear. There is a growing evidence of the role of CGIs methylation in the regulation of microRNAs (Han et al. 2007). In particular, it has been demonstrated that 80% of the promoters of ''intergenic'' microRNAs contain CGIs. In addition, these regulatory regions show signals of evolutionary conservation (Wang et al. 2010). We also cannot exclude that some CGIs categorized as intergenic, may be related to yet unidentified genes.

Bock et al. (Bock et al. 2006) developed a computational epigenetics approach to discriminate between CpG islands that are prone to methylation from those that remain unmethylated on the basis of a set of 1,184 DNA attributes. One of these attributes was the evolutionary conservation that the authors found to be uncorrelated with CpG island methylation. It should be noted that in this study (published in 2006) only CGIs on chromosome 21 were analyzed. Further, the methods to evaluate evolutionary conservation and for the statistical analysis are not the same that we used.

Our study has some limit. The most important one is the estimation of CGIs methylation. For each CGI we have data only on a limited number of CpGs, and from their methylation values we estimated the total CGI methylation. It should be noted that the dataset that we used is the largest genome-wide dataset available and that, in any case, this could be considered a systematic error that could cause a general noise only.

Another limit is that we analyzed the DNA methylation only. Epigenetic control of transcription involves a complex network of signals, including transcription factors, noncoding RNAs, DNA methylation, and histone modifications (Bonasio et al. 2010). In this study we looked only to a part of these mechanisms. Further studies are needed to analyze the other component of this machinery.

Another possible limit concerns the method used to define regions under selective pressure. Other methods have been described and our choice could not be exhaustive. A final caveat concerns possible cell-culture induced DNA methylation. It is well established that in vitro culture can cause changes in epigenetic marking of the genome (Bork et al. 2010; Saferali et al. 2010), probably due to the adaptation of the cells to the in vitro conditions. Therefore it should be underlined that, concerning DNA methylation, cell lines could be not representative of their relative primary tissues.

In conclusion, in this paper we demonstrated, in several cell lines, that CpG islands in regions showing signatures of selective pressure are undermethylated in comparison with the other regions of the genome. Additionally, by analyzing SNP frequency in CpG islands, we demonstrated that CpG islands in regions under selective pressure show lower genetic variation among individuals.

Materials and Methods

Data and evolutionary scores

All the data and the scores that we used were downloaded from annotation tracks in the UCSC Genome Browser (Sanborn et al. 2011). A brief description is provided below. Further and more detailed information about the dataset used can be found at http://genome.ucsc.edu/.

CpG island coordinates

CGIs genomic coordinates were obtained from the UCSC GB CpgIslandExt track. In this track CpG islands were predicted by searching the sequence one base at a time, scoring each dinucleotide (+17 for CG and 21 for others) and identifying maximally scoring segments. In this dataset, to define a CpG island the following criteria were used: i) to have a GC content of 50% or greater, ii) to have a length greater than 200 bp, and iii) to show a ratio greater than 0.6 of observed number of CG dinucleotides to the expected number, calculated on the basis of the number of Gs and Cs in the segment under analysis.

DNA methylation data

Methylation profiles from each cell sample were downloaded from the UCSC GB HAIB Methyl RRBS Track. These tables report the percentage of DNA molecules that show cytosine methylation at specific CpG dinucleotides in several cell lines. To obtain these data, researchers belonging the ENCODE Consortium assayed DNA methylation at CpG sites with a modified version of Reduced Representation Bisulfite Sequencing (Meissner et al. 2008). We used data from 25 cell lines, which were the first ones to come out from the moratorium period (expiration of moratorium period =2011-04-13). The data set contains, for each cell line at least two replicas each containing, on average, about 1.5 million of CpG methylation values. To exclude unreliable data, only methylation signals identified by a number of reads ≥10 were used for further analyses. After this filtering, we computed, for each CpG the mean value between two replicas, obtaining methylation values of genomic CpGs per cell line in the range (5–8) 10^5. We next selected methylation values of CpG dinucleotides in CGIs, filtering them by the CpG Islands track of UCSC-GB. The final CGI methylation value was obtained by calculating the mean methylation of all CpGs contained in the CGI.

Integrated haplotype score (iHS)

The normalized his scores were obtained from UCSC Genome Browser ‘‘HGDP iHS’’ track. The per-continent integrated haplotype score (iHS) (Voight et al. 2006) is a measure of recent positive selection. The scores present in the UCSC Genome Browser were calculated using SNPs genotyped in 53 populations worldwide by the Human Genome Diversity Project in collaboration with the Centre d’Etude du Polymorphisme Humain (HGDP-CEPH).

Samples from 1043 individuals from different geographical regions were genotyped for 657000 SNPs at Stanford. The 53 populations were divided into seven continental groups: Africa (Bantu populations only), Middle East, Europe, South Asia, East Asia, Oceania and the Americas.

iHS was calculated for each population group and then normalizing the resulting unstandardized iHS scores in derived allele frequency bins as described in (Voight et al. 2006). Per-SNP iHS scores were smoothed in windows of 31 SNPs, centered on each SNP. The final score is 2log10 of the proportion of smoothed scores higher than each SNP’s smoothed score.

We converted genome coordinates from assembly NCBI36/hg18 to assembly GRCh37/hg19 by using Batch Coordinate Conversion (liftOver) utility (UCSC Genome Browser). We scanned normalized iHS scores across the whole genome and selected the genomic intervals where iHS values ≥2. Once detected such compact regions, we extended their boundaries to the nearest loci where iHS was exactly vanishing.

Selective Sweep Scan: 5% Smallest S scores

Green et al. identified polymorphic sites among five modern human genomes and determined ancestral or derived state of each single SNP (Green et al. 2010). The human allele states were used to estimate an expected number of derived alleles in Neanderthal in the 100000-base window around each SNP. The measure called S score compare the observed number of Neanderthal alleles in each window to the expected number. An S score significantly less than zero indicates an increase of human-derived alleles not found in Neanderthal, suggesting positive selection in the human lineage since divergence from Neanderthals. Regions with S scores in the lowest 5% (strongest negative scores, ''5% Lowest S'' track of UCSC Genome Browser) were used in our analyses.

Conserved Elements

Conserved elements were downloaded from the UCSC GB Conservation (cons46way) Track. In this track conserved elements were predicted using the methods phastCons and phyloP. Both phastCons and phyloP are phylogenetic methods that rely on a tree model containing the tree topology, branch lengths representing evolutionary distance at neutrally evolving sites, the background distribution of nucleotides, and a substitution rate matrix. Pairwise alignments with the human genome were generated for each species using blastz from repeat-masked genomic sequence. The conserved elements were predicted using 10 primate species. Primate species used are: *Homo sapiens* (reference species), *Pan troglodytes, Gorilla gorilla gorilla, Pongo pygmaeus abelii, Macaca mulatta, Papio hamadryas, Callithrix jacchus, Tarsier syrichta, Microcebus murinus, Otolemur garnettii.*

Statistical analysis

In order to test the null hypothesis that two distributions have the same means we use a ''bootstrapping approach''. In particular we take the mean of the smaller sample, hereafter denoted by m, and compare this value with the probability distribution of mean, p(m), obtained from a large number (10^4) of randomly sampled cohorts of the same size taken from the larger sample. Type I error to reject the null hypothesis even if it is true, denoted as ''Bootstrap p-value'' of the test, by definition is the sum of p(m) for m≥m. Since we have 10^4 cohorts of the larger sample the precision of our ''Bootstrap p-value'' is 10^{-4}, which is however small enough since we have fixed the threshold of statistical significance at 10^{-3}. All statistical analyses were carried out with Galaxy (http://galaxyproject.org/) and R ver. 2.10.1 (R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org/).

Chapter 3: CpG islands under selective pressure are enriched with H3K4me3, H3K27ac and H3K36me3 histone modifications

Introduction

CpG islands (CGIs) are unmethylated segments of a genome that have an increased level of CpG dinucleotides and a high GC content (Bird et al. 1985; Illingworth and Bird 2009). In the human genome, most CGIs are either inside or close to the promoter regions of genes (Larsen et al. 1992). Typically these CGIs occur at or close to transcription start sites (TSSs) (Illingworth et al. 2010). It is well established that CpG sites in promoter CGIs are undermethylated in expressed genes, while hypermethylation of promoter CpG sites is associated with gene silencing (Jones 2012). Others CGIs that are distant from known TSSs have been found in intergenic, 3' and intragenic regions (Medvedeva et al. 2010).

There is an extensive literature demonstrating that structural modifications to chromatin, along with CGI methylation, contribute to the functional output of related genes (Blackledge and Klose 2011). The N-terminal tails of histone proteins can be modified covalently by small molecules (for example, phosphorylation, acetylation, methylation) and by macromolecules (for example, ubiquitination, sumoylation etc.). The precise environment of the CGI chromatin that controls gene regulation is not definitively established. The general understanding is that by altering the state of the CGI chromatin, histone modification can regulate access of the transcription machinery to particular DNA sequences (Li et al. 2007). Of all the possible histone modifications, methylation of the lysine or arginine residues has received the main attention. These modifications can activate or repress the associated genes depending on which lysine or arginine residues are methylated (Kouzarides 2007). Methylation of histone H3 at lysine 9 (H3K9) or lysine 27 (H3K27) is considered to be a repressive mark (Kouzarides 2007). In contrast, H3K4me3, perhaps the best established epigenetic marker, is robustly associated with activation of transcription (Kouzarides 2007). In mammals, the trimethylation of H3k4 can be catalyzed by different histone methyltransferases, such as MLL1 or ASH1L (Dou et al. 2006; Gregory et al. 2007). The majority of H3K4me3 sites overlap with the 5' ends of annotated human genes (Guenther et al. 2007) and several studies have reported the inverse correlation between two epigenetic marks, DNA methylation and H3K4me3 (Okitsu and Hsieh 2007; Balasubramanian et al. 2012). The H3K4me3 mark also plays a crucial role in mammalian development (Bernstein et al. 2006), and its alteration has been found to be associated with cancer and other diseases (Kaneda et al. 2009; Ke et al. 2009; Sandgren et al. 2010). In addition, both H3K27ac and H3K36me3, which are known as a promoter mark (Wang et al. 2008) and a gene body mark (Barski et al. 2007), respectively, are associated with transcriptional activation (Wang et al. 2008; Wagner and Carpenter 2012).

Alterations in gene regulation are thought to play an important role both in adaptation and evolution (Britten and Davidson 1969). A recent report proposed that differences in gene expression levels among primates are associated with the changes in H3K4me3 (Cain et al. 2011). Moreover, another recent study identified human-specific changes in H3K4me3 levels at TSSs and related regulatory sequences in comparison with chimpanzees and macaques (Shulha et al. 2012). Besides, the Encyclopedia of DNA Elements (ENCODE) project is studying different functional elements of human genome including regions of histone modifications. In particular they assayed chromosomal locations for 12 histone modification in 46 different cell types (Dunham et al. 2012). In a previous study

(Cocozza et al. 2011), we demonstrated that CGIs under selective pressure are hypomethylated compared to the CGIs in other regions of the genome. In this study, we explored the relationship between selective pressure signatures of and histone modification (H3K4me3, H3K27ac and H3K36me3) enrichment in CGIs. We used the genome-wide histone modification data of thirteen human cell lines produced by the ENCODE consortium (Celniker et al. 2009; Ernst et al. 2011). To define regions under selective pressure we used three distinct methods (Cocozza et al. 2011) that are able to detect both recent and ancient selective pressure events (Nielsen 2005).

Results

We analyzed thirteen cell lines from the ENCODE/Broad Institute, derived from nine normal and four cancer tissues, respectively. A list of features for each considered cell line is presented in Supplementary table 3.1. For each cell line, we downloaded histone modification data for H3K4me3, H3K27ac and H3K36me3 marks. We used the "Peaks Signal" (PS), representing regions of statically significant enrichment of a specific histone modification (see Materials and Methods). We downloaded genomic coordinates of 27718 unique CGIs defined according to criteria described in the University of California Santa Cruz Genome Browser (UCSC GB) (http://genome.ucsc.edu/) (see Materials and Methods). For each cell line, we estimated the number of CGIs containing at least one PS of histone modification and found, on average, 15478, 11903 and 10182 CGIs containing PSs of H3K4me3, H3K27ac and H3K36me3, respectively.

To identify genomic regions that may have undergone selective pressure we used three different approaches that are sensitive to selective pressure events that occurred in distinct evolutionary epochs.

The first method uses the per-continent "integrated Haplotype Score" (iHS) (Voight et al. 2006) and marks recent positive selection (see Materials and Methods). Using the iHS we identified 586 genomic regions that have putatively undergone recent selective pressure. We denoted these regions as ''high iHS regions'' (HIRs). Within the HIRs regions we found 2545 CGIs.

The second approach is based on a comparison between *Homo sapiens* and *Neanderthal* genomes (see Materials and Methods). The selective sweep scan score (S score) was used to identify regions of the human genome with a strong signal for depletion of Neanderthal-derived alleles. This score, when negative, may indicate an episode of positive selection in early humans (Green et al. 2010). We found 212 genomic regions with a significant negative score (5% lowest S regions, hereafter denoted as 5LSRs) containing 348 CGIs.

In the third approach, we looked for sequences that were conserved across ten primate genomes. These sequences are the so-called "Conserved Elements" (CEs) (see Materials and Methods) and they mark ancient selective pressure events. We downloaded 725627 CEs and used them to search for CGIs that contain CEs (Siepel et al. 2005). We identified 13288 unique CGIs that contained at least one CE.

We computed the fraction of CGIs containing histone modification marks that show signatures of natural selection (HIRs, CEs and 5LSRs), and compared it with an analogous quantity computed for CGIs shown to have no signals of selective pressure. The presence of a possible

enrichment/diminishment, defined as the ratio of the percentages of the above two groups, was assessed by means of a hypergeometric test (see Material and Methods).

Overall Analysis

We found a significant enrichment of H3K4me3 and H3K27ac markers for all three signatures of selection in almost all cell lines (Figure 3.1, Figure 3.2 and Supplementary table 3.2) while for H3K36me3 the enrichment reached significance only for the CE signature (Figure 3.3 and Table 3.2).

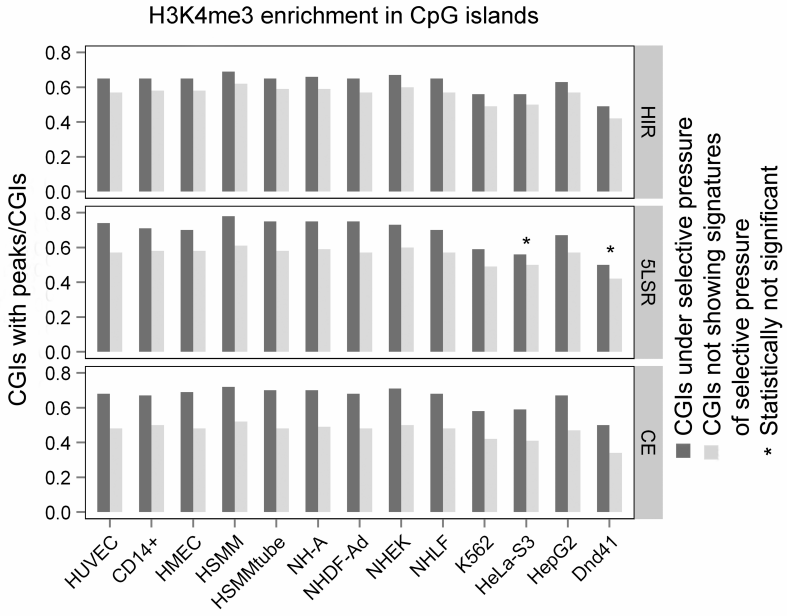

Figure 3.1: Enrichment of H3K4me3 modification in CpG islands under selective pressure. Black bars represent the fraction of CGIs containing histone modification marks within regions that show signatures of natural selection (HIRs, CEs and 5LSRs). Grey bars represent the fraction of CGIs containing histone modification marks within regions that do not show signatures of selective events. The X-axis indicates the analyzed cell lines. An asterisk (*) above a bar indicates a statistically non-significant difference.

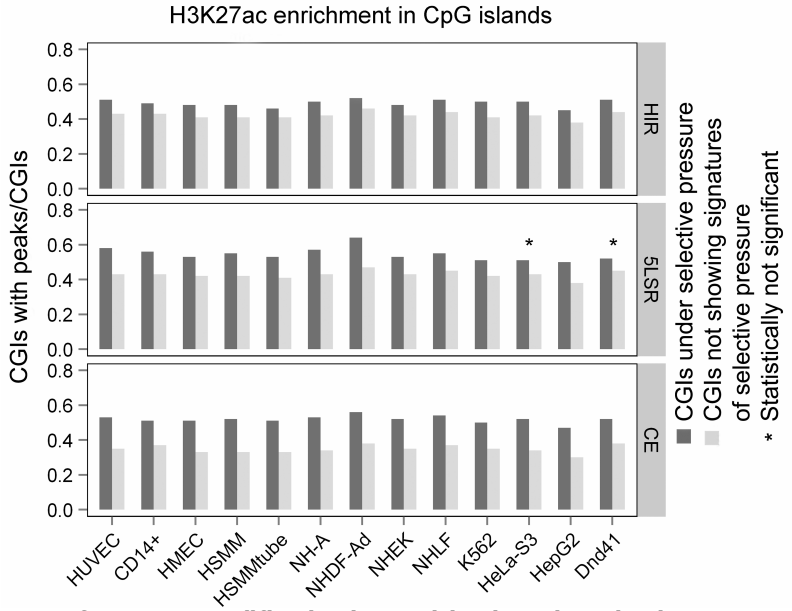

Figure 3.2: Enrichment of H3K27ac modification in CpG islands under selective pressure. Same notation as in Figure 3.1.

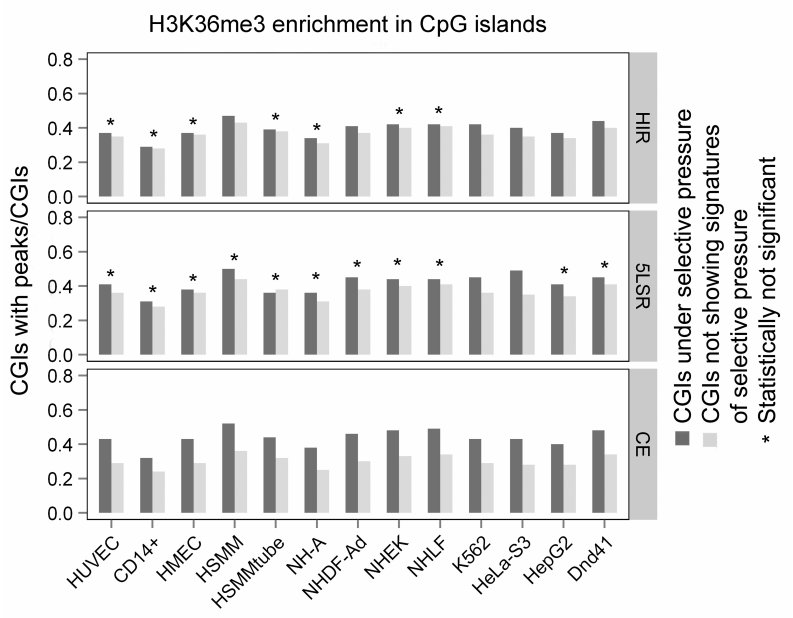

Figure 3.3: Enrichment of H3K36me3 modification in CpG islands under selective pressure. Same notation as in Figure 3.1.

Position Analysis

We investigated whether or not these differences were dependent on the position of the CGIs in various genomic regions. To do this we followed the same approach described by Medvedeva et al. (Medvedeva et al. 2010) dividing the CGIs into four groups according to their positions with respect to genes: at the 5' end of a gene, in the intragenic region, at the 3' end of a gene, and in the intergenic region. Results of this analysis are presented in the Supplementary table 3.3 and summarized below.

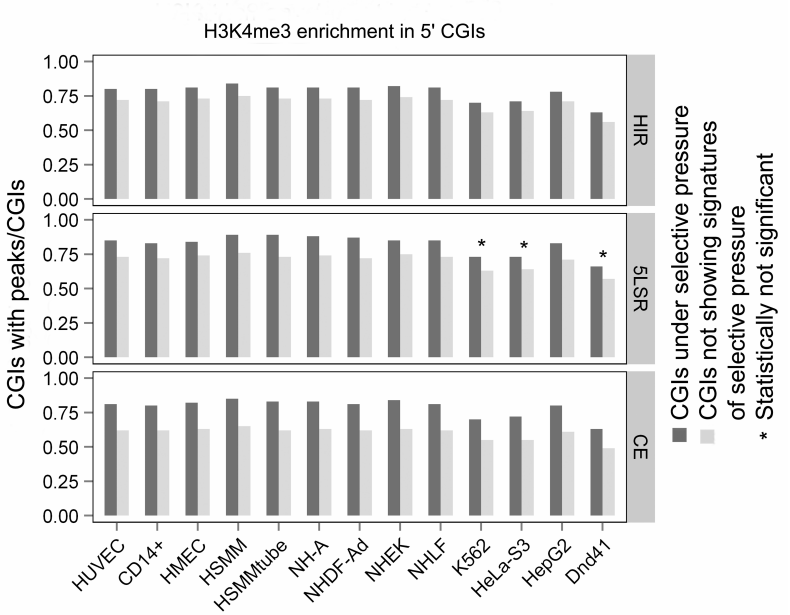

Figure 3.4: Enrichment of H3K4me3 modification in 5' CpG islands under selective pressure. Same notation as in Figure 3.1.

Analysis of 5' CGIs demonstrated the same significant enrichment pattern as seen in the overall analysis with significant enrichment of H3K4me3 and H3K27ac (Figure 3.4 and Figure 3.5), in almost all cell lines for all signatures of selection, and significant enrichment of H3K36me3 in all cell lines for the CE signature only (Figure 3.6).

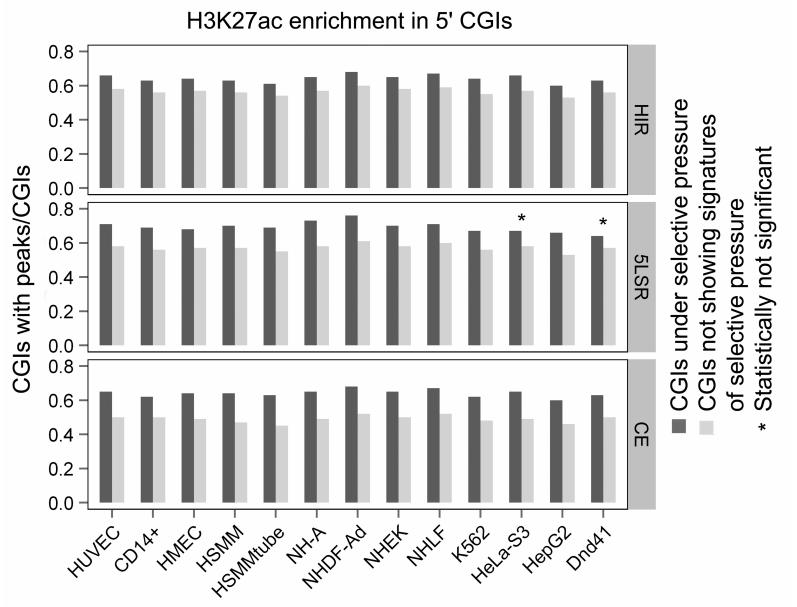

Figure 3.5: Enrichment of H3K27ac modification in 5' CpG islands under selective pressure. Same notation as in Figure 3.1.

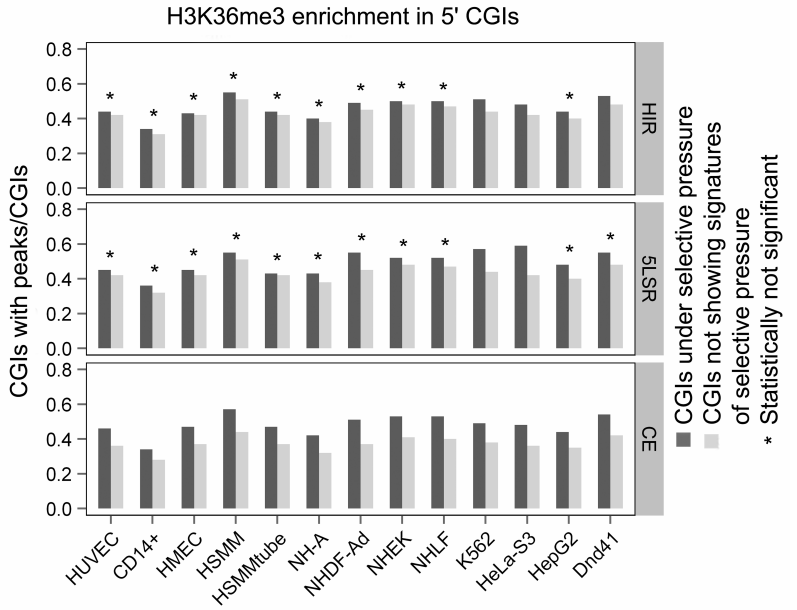

Figure 3.6: Enrichment of H3K36me3 modification in 5' CpG islands under selective pressure. Same notation as Figure 3.1.

Both intragenic and 3' CGIs were significantly enriched for H3K36me3 in the majority of cells lines (nine and ten out of the thirteen, respectively) for the CE signature (Figure 3.7 and Figure 3.8), while

analysis of the other markers did not reach significance in almost all other cases (Figure 3.9, Figure 3.10, Figure 3.11 and Figure 3.12)

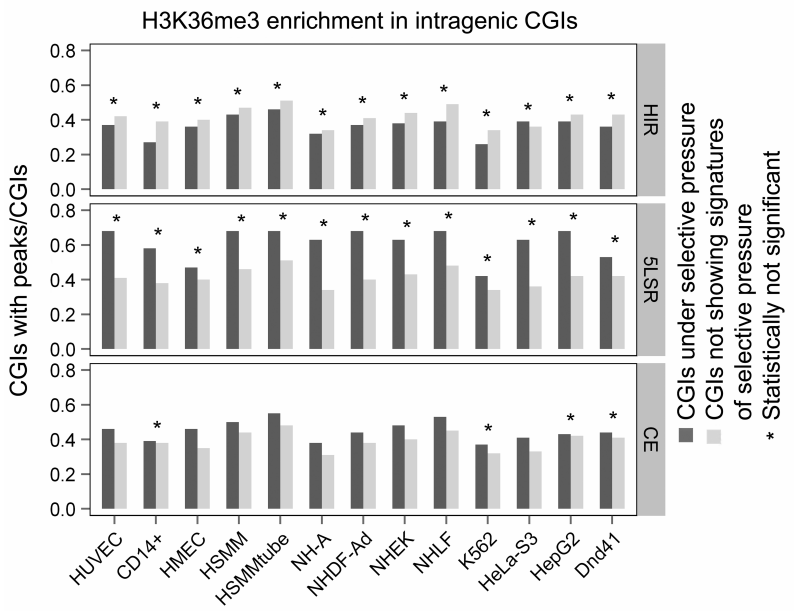

Figure 3.7: **Enrichment of H3K36me3 modification in intragenic CpG islands under selective pressure.** Same notation as Figure 3.1.

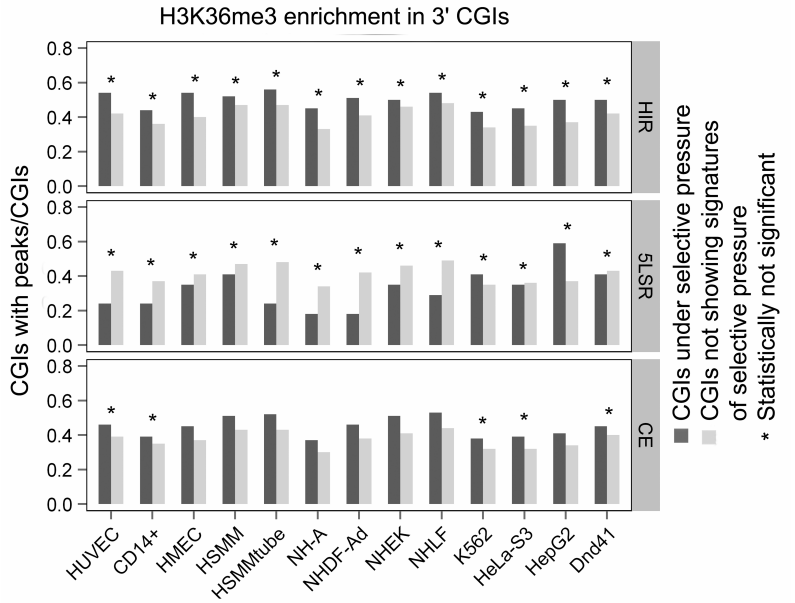

Figure 3.8: **Enrichment of H3K36me3 modification in intragenic CpG islands under selective pressure.** Same notation as Figure 3.1.

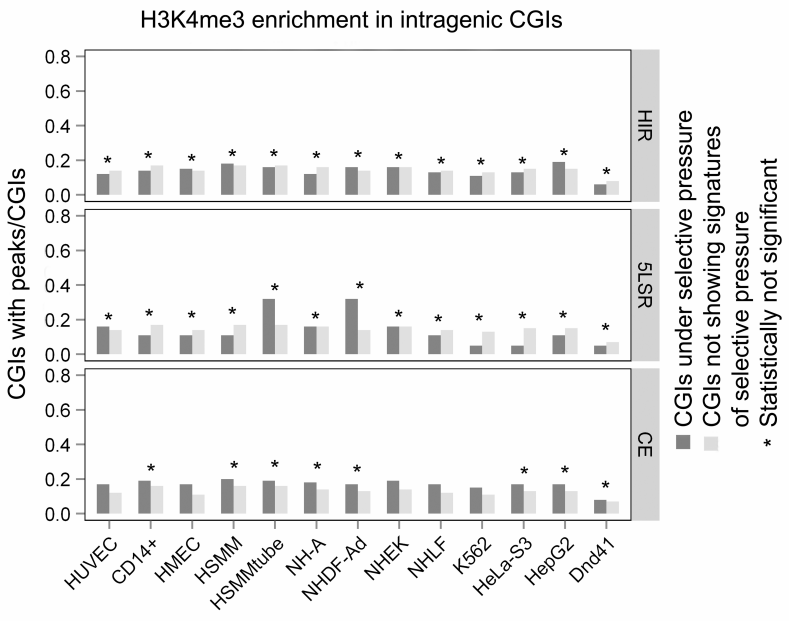

Figure 3.9: Enrichment of H3K4me3 modification in intragenic CpG islands under selective pressure. Same notation as Figure 3.1.

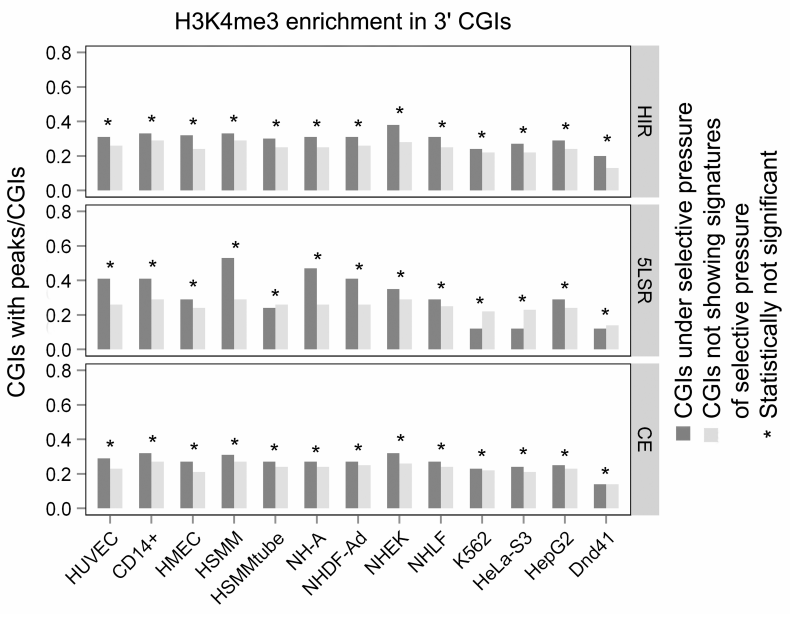

Figure 3.10: Enrichment of H3K4me3 modification in 3' CpG islands under selective pressure. Same notation as Figure 3.1.

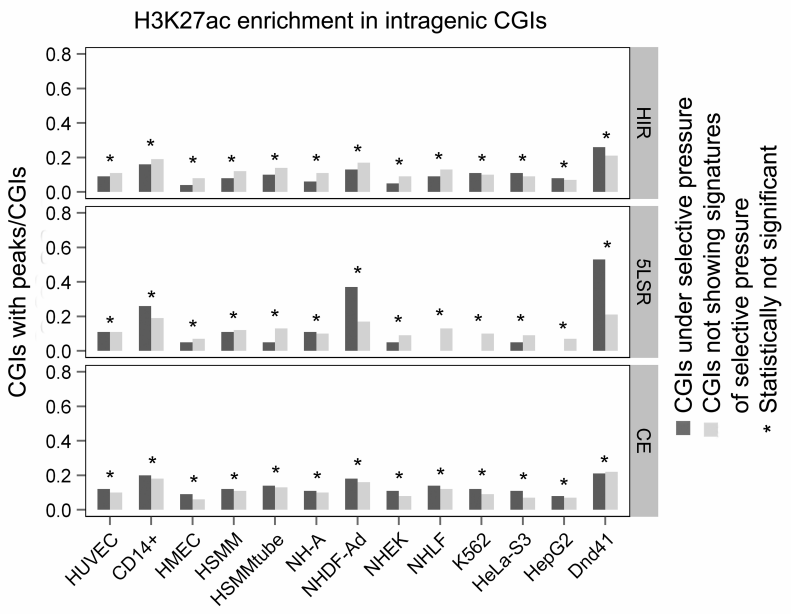

Figure 3.11: Enrichment of H3K27ac modification in intragenic CpG islands under selective pressure. Same notation as Figure 3.1.

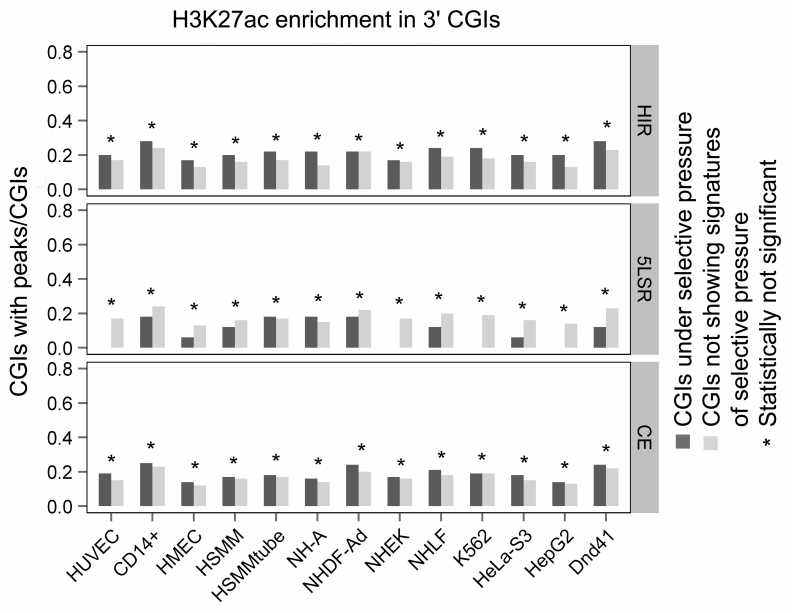

Figure 3.12: Enrichment of H3K27ac modification in 3' CpG islands under selective pressure. Same notation as Figure 3.1.

Finally, regarding intergenic CGIs we found a significant enrichment in all cell lines for all considered markers for the CE signature and in twelve out of thirteen cell lines for H3K4me3 in the HIR signature (Figure 3.13, Figure 3.14 and Figure 3.15).

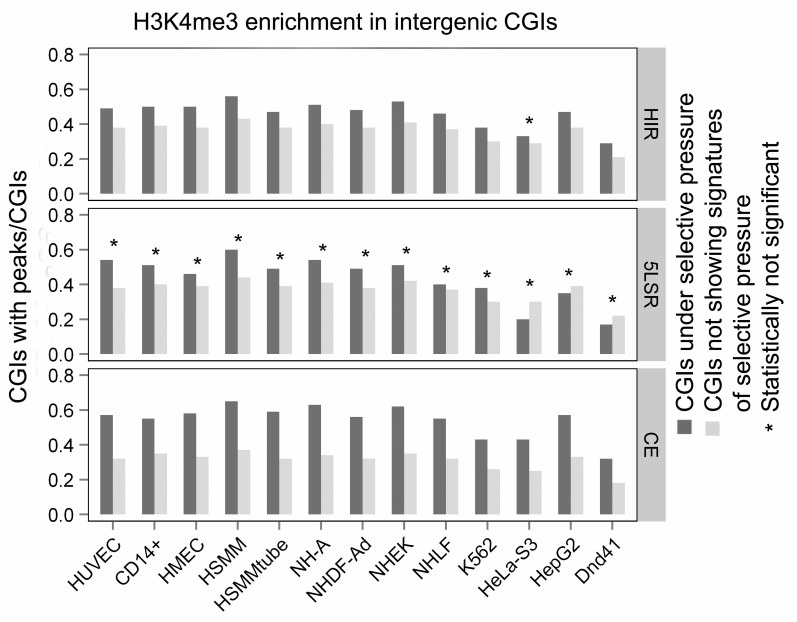

Figure 3.13: Enrichment of H3K4me3 modification in intergenic CpG islands under selective pressure. Same notation as Figure 3.1.

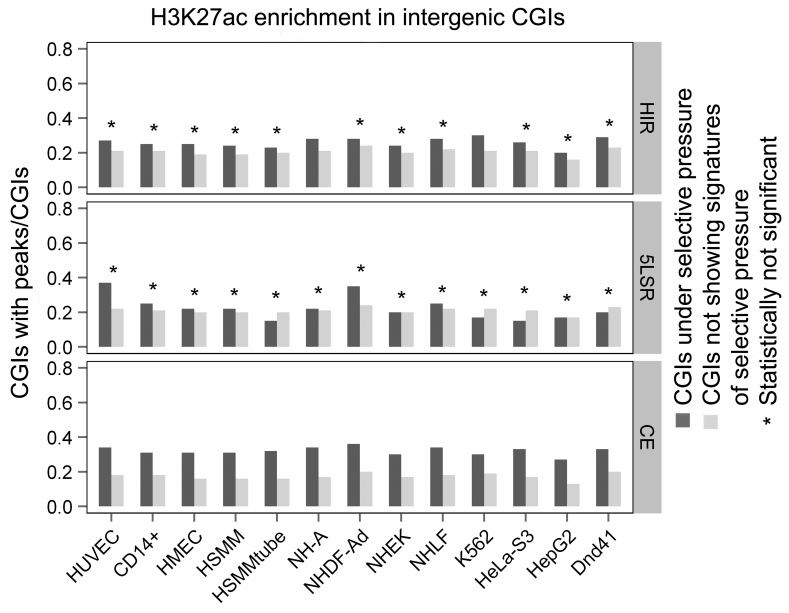

Figure 3.14: Enrichment of H3K27ac modification in intergenic CpG islands under relective pressure. Same notation as Figure 3.1.

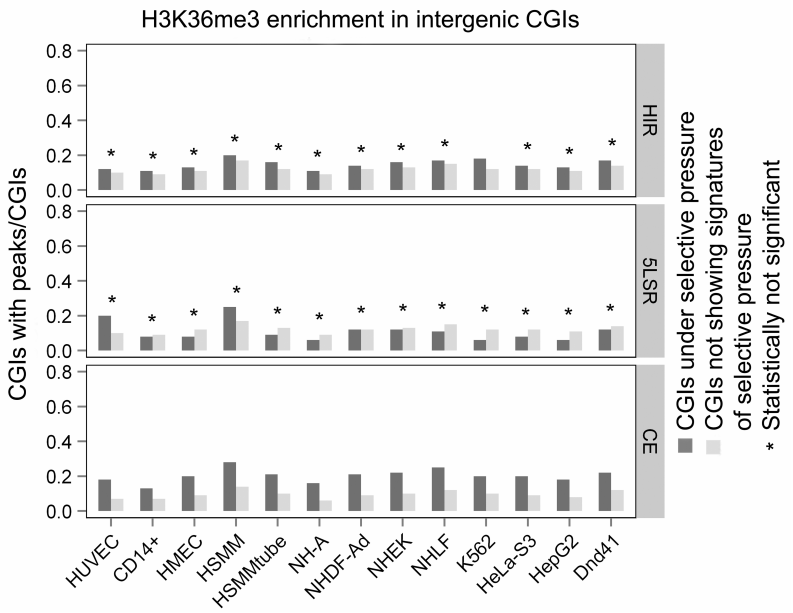

Figure 3.15: Enrichment of H3K36me3 modification in intergenic CpG islands under selective pressure. Same notation as Figure 3.1.

Evolutionary forces analysis

Two major evolutionary forces result in enriched CpG content: one is based on low levels of DNA methylation and, consequentially, deamination; and the other is biased gene conversion (BGC), which acts to repair TG mismatches caused by the deamination of methyl-cytosine (Cohen et al. 2011). According to the role that these two forces play in CGI maintenance, CGIs can be classified as hypo-deaminated CGIs or BGC CGIs. We examined whether or not the relationship that we found between selective pressure and histone mark enrichment was present in both classes.

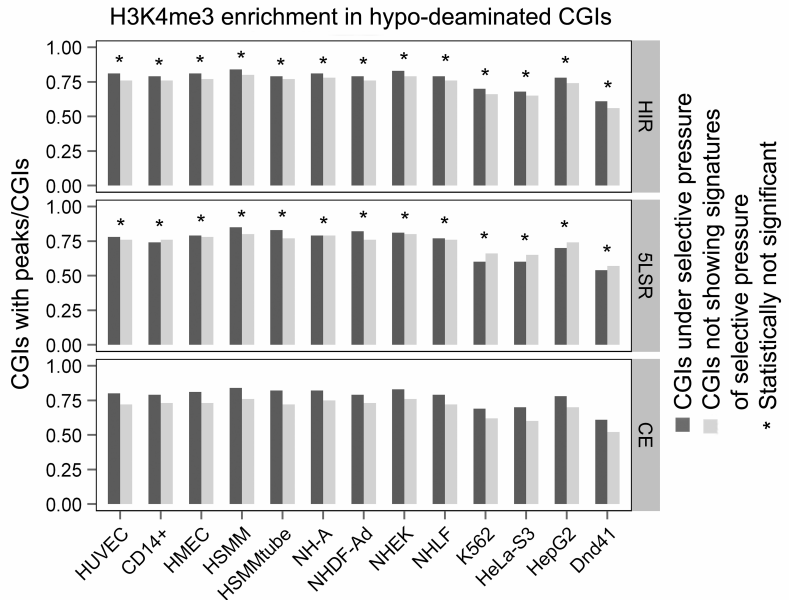

Figure 3.16: Enrichment of H3K4me3 modification in hypo-deaminated CpG islands under selective pressure. Same notation as Figure 3.1.

We found that both hypo-deaminated and BCG CGIs showed an enrichment of all markers in the CE signature in all cell lines, while only BGG CGIs showed significant enrichment of H3K4me3 and H3K27ac in the HIR signature in all cell lines (Figure 3.16, Figure 3.17, Figure 3.18, Figure 3.19, Figure 3.20, Figure 3.21 and Supplementary table 3.4).

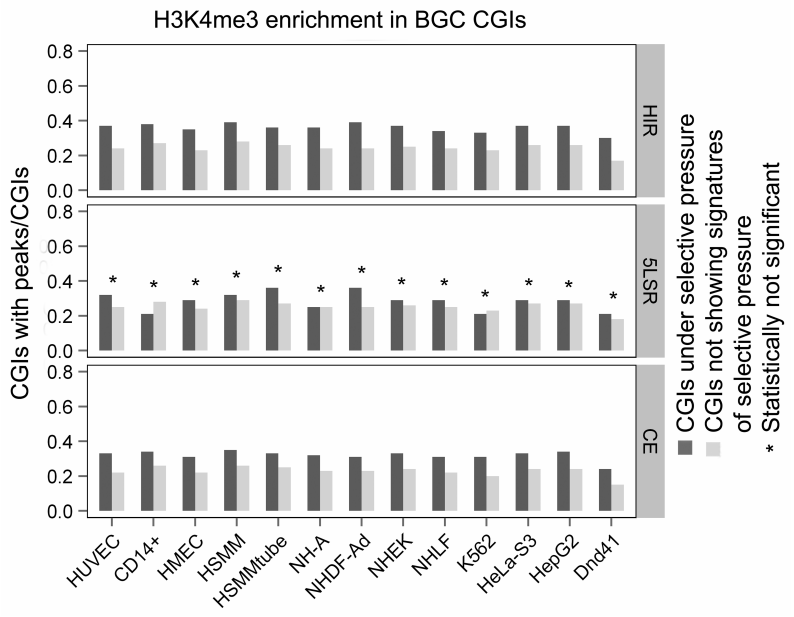

Figure 3.17: Enrichment of H3K4me3 modification in BGC CpG islands under selective pressure. Same notation as Figure 3.1.

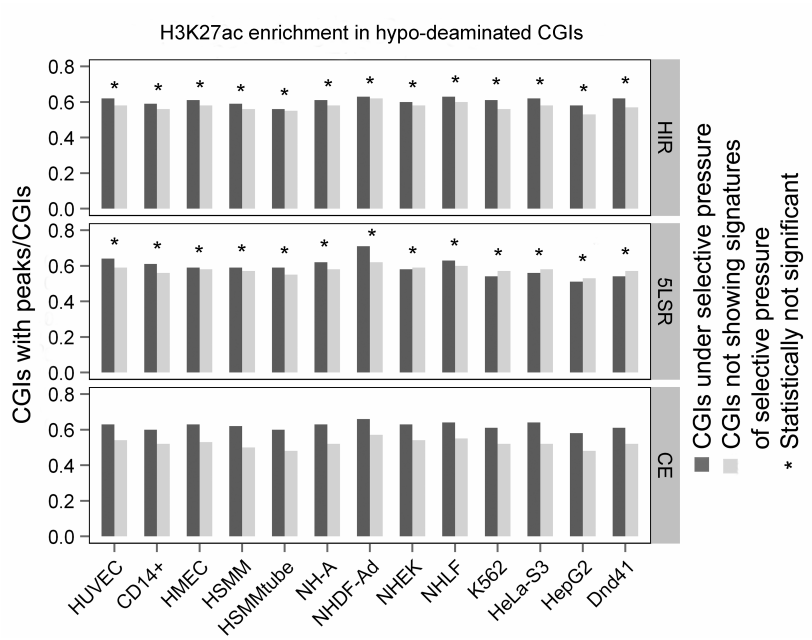

Figure 3.18: Enrichment of H3K27ac modification in hypo-deaminated CpG islands under selective pressure. Same notation as Figure 3.1.

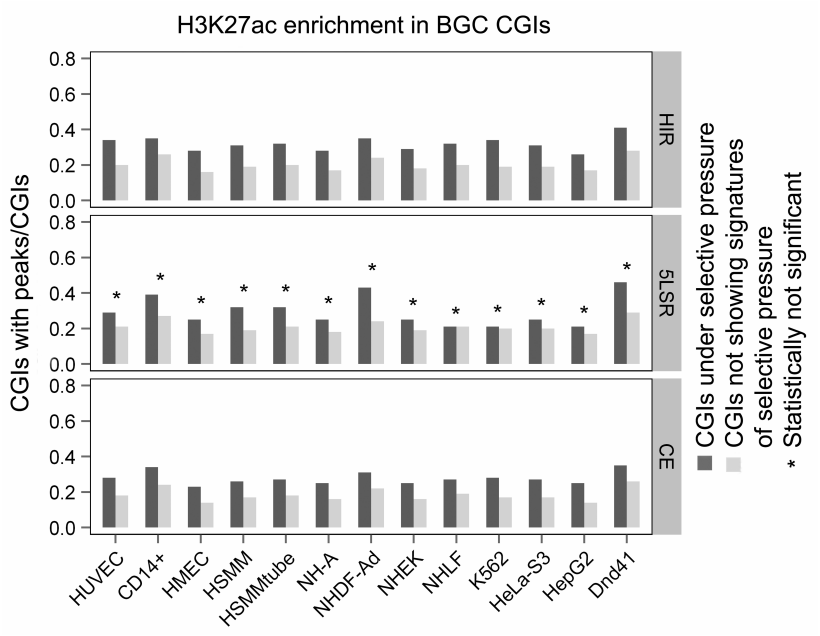

Figure 3.19: Enrichment of H3K27ac modification in BGC CpG islands under selective pressure. Same notation as Figure 3.1.

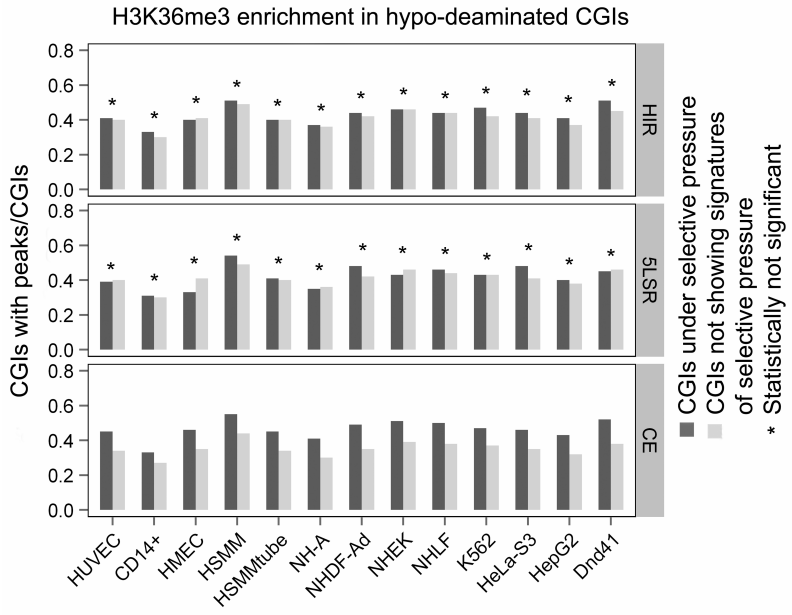

Figure 3.20: Enrichment of H3K36me3 modification in hypo-deaminated CpG islands under selective pressure. Same notation as Figure 3.1.

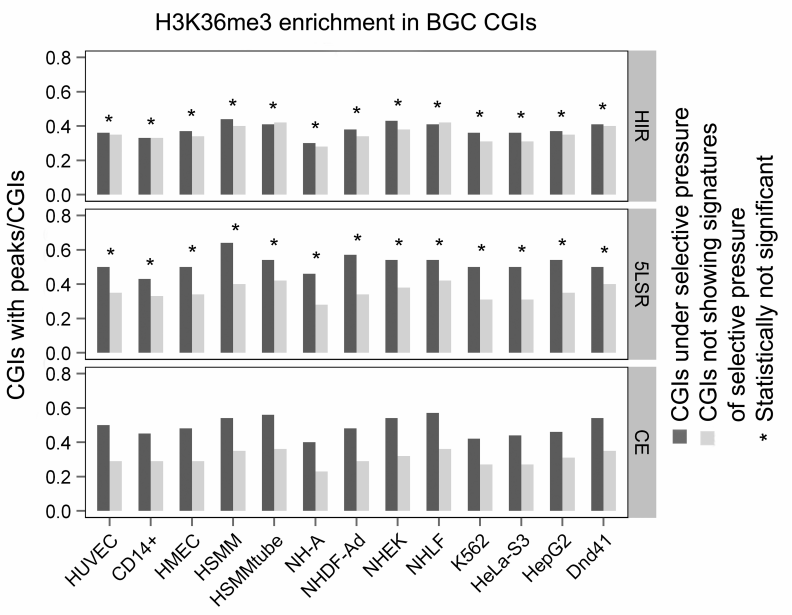

Figure 3.21: Enrichment of H3K36me3 modification in BGC CpG islands under selective pressure. Same notation as Figure 3.1.

Discussion

In this study, we investigated the hypothesis that CGIs under selective pressure are enriched with histone modifications that are associated with gene activation. To do this, we analyzed data from thirteen human cell lines for three well-known histone modifications (H3K4me3, H3K27ac and H3K36me3) to explore their relationship with both recent and ancient events of selective pressure.

H3K4me3 and H3K27ac are epigenetic marks that are generally associated with gene activation (Kouzarides 2007; Wang et al. 2008) while H3K36me3 is associated with transcriptional elongation (Li et al. 2007). Moreover, H3K4me3 and H3K27ac are evolutionarily conserved among species (Woo and Li 2012) and negatively correlated with DNA methylation (Okitsu and Hsieh 2007; Thomson et al. 2010; Bell et al. 2011; Balasubramanian et al. 2012). Also H3K36me3 in exons is found to be conserved between human and mouse (Kolasinska-Zwierz et al. 2009).

Using the entire set of human CGIs we found that the CGIs associated with signatures of selective pressure were significantly enriched with H3K4me3 and H3K27ac in almost all considered cell lines. H3K36me3, on the other hand, showed a significant enrichment in global CGIs only in CE regions: this could be due to a small sample size effect (Table S2). These findings support a previous study in which we found that CGIs located in regions under selective pressure are more protected from DNA methylation compared the CGIs in other genomic regions (Cocozza et al. 2011).

When we divided CGIs according to their positions with respect to the genes, we found that the statistical differences between CGIs with and without signatures of selective pressure were clearest for CGIs located in the 5' regions for both H3K4me3 and H3K27ac. This result is intriguing in the light

of the well established evidence that CGIs at the 5' ends of genes are involved mainly in the control of gene expression (Ruthenburg et al. 2007). It is also possible that the small sample size led to a lack of statistical confidence in the results for CGIs in other positions. We noticed a different behavior for H3K36me3. H3K36me3 was the only mark to be enriched in 3' and intragenic CGIs in CE regions for majority of cell lines; this finding is again intriguing considering that H3K36me3 is reported to be a gene body mark (Barski et al. 2007; Li et al. 2007).

Two different evolutionary processes, namely hypo-deamination and BCG, are involved in the generation and maintenance of CGIs (Cohen et al. 2011). The majority of hypo-deaminated CGIs are usually unmethylated while most BGC CGIs are constitutively methylated and clustered in subtelomeric regions. We found H3K4me3, H3K27ac and H3K36me3 enrichment in CGIs in CE regions, independently of the evolutionary process involved in their generation. Since CGIs belonging to these two groups differ in their DNA methylation levels, our finding seems to suggest that the difference we found was quite independent of the DNA methylation status.

The impact of natural selection on functional elements in human genome is also addressed in the last report from ENCODE project (Dunham et al. 2012). In that case the authors focused their attention mainly on the relationship between negative selection and a subset of functional elements but they did not specifically address histone modifications. Positive selection, on the other hand, was addressed in a recent work by Vernot et al. (Vernot et al. 2012) who studied the impact of this kind of selective pressure on DNase I peaks.

It has been hypothesized that CGIs are fundamental regulatory structures that have evolved under selection in genomes where DNA methylation plays a regulatory role (Deaton and Bird 2011; Shulha et al. 2012). In particular, CGIs act as a platform where chromatin modifications and additional signaling help to define the functional output of the respective genes. The present study concerning H3K4me3, H3K27ac and H3K36me3 enrichment in CGIs under selective pressure, supplements a previous study by Cocozza et al (Cocozza et al. 2011).

To our knowledge, the present study is the first report addressing the relationship between histone modifications and natural selection and the overall framework emerging from our analyses support the hypothesis that CGIs that have experienced selection could be characterized by distinct epigenetic signatures.

Materials and methods

Histone modification data

The histone modification (H3K4me3, H3K27ac and H3K36me3) data for thirteen human cell lines (HUVEC, Monocytes-CD14+-RO01746 (CD14+), HMEC, HSMM, HSMMtube, NH-A, NHDF-Ad, NHEK, NHLF, K562, HeLa-S3, HepG2 and Dnd41) were downloaded from the "Broad histone" track of the UCSC GB. This track contains genome-wide histone modification data of different cell lines, generated using ChIP-seq high-throughput sequencing as a part of the ENCODE project (Ernst et al. 2011). In this study, we used the "Peaks Signal" (PS), which identifies discrete intervals of ChIP-seq fragment enrichment. In particular, we considered the CGIs in our sample that contained at least one PS.

CpG islands

UCSC CGIs: CGIs coordinates were downloaded from the "CpgIslandExt" track of the UCSC GB (http://genome.ucsc.edu/). The CGIs in this track were predicted by searching the human genome assembly (GRCh37/hg19) sequence, scoring each dinucleotide and identifying maximally scoring segments. In this dataset, a CpG island was defined according to the following criteria: i) GC content of 50% or greater, ii) length of at least 200 bp, and iii) observed CpG / expected CpG ratio greater than 0.6. The CGI set that we obtained consisted of 27718 CGIs (this excluded the CGIs in the data related to the alternative haplotype sequences).

5', intragenic, 3', and intergenic CGIs: We used the classification system that was described previously by Medvedeva et al. (Medvedeva et al. 2010) in which the CGIs were classified according to their locations. Thus, the CGIs were classified into four classes:

1) 5' CGIs - located in the 5' flank region (3 kb upstream the TSS), the 5' UTR-exon, the 5' UTR- intron, the initial coding exon or the initial intron.
2) Intragenic CGIs are located in the internal exons and introns.
3) 3' CGIs are located in the final exon, the final introns, the 3' UTR-exon or in the 3' UTR-intron.
4) Intergenic CGIs are located at least 3 kb upstream or downstream from any known gene.

Hypo-deaminated and biased gene conversion (BGC) CGIs: Two sets of CGIs were described by Cohen et al. (Cohen et al. 2011) using a new parameter-rich evolutionary model in combination with high resolution DNA methylation data to study the origin of the CpG repertoire in primate genomes (marmoset, rhesus, orangutan, chimp and human). Following a clustering analysis, they observed that most CGIs were constitutively unmethylated and underwent slow C-to-T deamination. They denoted this group as hypo-deaminated CGIs. In contrast, another class of CGI was constitutively methylated with a rapid deamination rate and was termed as BGC CGIs. For our analysis, we considered the 9091 hypo-deaminated and 4782 BGC CGIs from the UCSC CGIs sample.

Integrated haplotype score (iHS)

The iHS belongs to the Extended Haplotype Homozygosity statistic "family" (Sabeti et al. 2002) and is a marker of recent positive selection (Voight et al. 2006). The iHS measures the decay of identity, as a function of distance, of haplotypes that carry a specified ''core'' allele. We downloaded the iHS normalized values from the ''HGDP iHS'' track of the UCSC GB. The scores were calculated using SNPs genotyped in 1043 individual taken from 53 populations worldwide by the Human Genome Diversity Project in collaboration with the Centre d'Etude du Polymorphisme Humain (HGDP-CEPH). The 53 populations were divided into seven continental groups: Africa (Bantu populations only), Middle East,

Europe, South Asia, East Asia, Oceania and the Americas. For each population group, the iHS was calculated and then normalized (Voight et al. 2006). Per-SNP iHSs were smoothed in windows of 31 SNPs, centered on each SNP. The final score is -log10 of the proportion of smoothed scores higher than each SNP's smoothed score. For our analysis, we used the Batch Coordinate Conversion (liftOver) utility (UCSC GB) to convert the genome coordinates from assembly NCBI36/hg18 to assembly GRCh37/hg19. We scanned the normalized iHSs across the whole genome and selected the genomic intervals where the iHS was ≥ 2. After these regions were identified, we extended their boundaries to the nearest loci where the iHS exactly vanished.

Selective sweep scan (S): the 5% lowest S scores (5LSR)

The S score is based on a comparison between *Homo sapiens* DNA and Neanderthal DNA (Green et al. 2010). We downloaded the regions with S scores from the ''5% Lowest S'' track of the UCSC GB and denoted them as "5LSRs" (5% lowest S regions). Green et. al. identified polymorphic sites among five modern human genomes and determined the ancestral or derived state of each SNP (Green et al. 2010). The states of the human alleles were used to estimate the expected number of derived alleles in Neanderthal in a 100000-base window around each SNP. The S scores were used to compare the observed number of Neanderthal alleles to the expected number in each window. A positive S score indicates more derived alleles in Neanderthal than expected given the frequency of derived alleles in human; a negative S score, on the other hand, indicates fewer derived alleles in Neanderthal, which might suggest positive selection in the human lineage after divergence from Neanderthal and before divergence in human populations. The 5LSRs represent the regions in the 5% lower percentile of the S score.

Conserved elements (CEs)

CEs are sequences in the genome that are conserved across species (Pollard et al. 2010). Conserved regions have a reduced rate of evolution compared to the expected rate under neutral drift. The CEs used in this study were downloaded from the "Conservation (cons46way)" track of the UCSC GB. This track shows measurements of evolutionary conserved elements using two phylogenetic methods, phastCons and phyloP. The CEs used in this study were predicted using ten primates, *Homo sapiens* (reference species), *Pan troglodytes, Gorilla gorilla*, *Pongo pygmaeus abelii*, *Macaca mulatta*, *Papio hamadryas*, *Callithrix jacchus*, *Tarsier syrichta*, *Microcebus murinus* and *Otolemur garnettii*.

Statistical analysis

We used a hypergeometric-based approach to test the null hypothesis that the possible enrichment of H3K4me3, H3k27ac and H3K36me3 is independent of the presence of signals of natural selection. In particular we considered: *k,* the observed number of CGIs containing both PSs and signatures of selective pressure, as the number of success in the sample; *n,* the number of CGIs characterized by signatures of selective pressure only, as the sample size; *M,* the total number of CGIs with PS, as the number of successes in the population; and *N,* the total number of CGIs, as the population size (see supplementary table 3.2-3.4). For statistical significance we set the threshold for the Bonferroni corrected p-value at 10^{-3}. All the statistical analyses were performed with R ver. 2.14.2 (R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org/).

General conclusion

Analyzing two different epigenetic marks such as DNA methylation and histone modifications (H3K4me3, H3K27ac and H3K36me3) of different human cell lines (including normal and cancer cell), we found that CGIs that experienced selective events are characterized by distinct epigenetic features. In particular, we found overall undermethylation of CGIs in the regions under selective pressure. On the other hand, we found in the same regions, enrichment of three histone marks, associated with active gene.

Further studies using other epigenetic marks could help to clarify the relation between epigenetic modifications and selective pressure in human genome.

Bibliography

Abraham, A.-L., M. Nagarajan, et al. (2012). "Genetic Modifiers of Chromatin Acetylation Antagonize the Reprogramming of Epi-Polymorphisms." PLoS Genetics **8**(9): e1002958.

Balasubramanian, D., B. Akhtar-Zaidi, et al. (2012). "H3K4me3 inversely correlates with DNA methylation at a large class of non-CpG-island containing start sites." Genome Med **4**(5): 47.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-37.

Bell, C. G., G. A. Wilson, et al. (2012). "Human-specific CpG "beacons" identify loci associated with human-specific traits and disease." Epigenetics **7**(10): 1188-99.

Bell, J. T., A. A. Pai, et al. (2011). "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines." Genome Biology **12**(1): R10.

Berger, S. L., T. Kouzarides, et al. (2009). "An operational definition of epigenetics." Genes & Development **23**(7): 781-3.

Bernstein, B. E., M. Kamal, et al. (2005). "Genomic maps and comparative analysis of histone modifications in human and mouse." Cell **120**(2): 169-181.

Bernstein, B. E., T. S. Mikkelsen, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell **125**(2): 315-326.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes & development **16**(1): 6-21.

Bird, A., M. Taggart, et al. (1985). "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA." Cell **40**(1): 91.

Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature **321**(6067): 209.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blackledge, N. P. and R. Klose (2011). "CpG island chromatin: a platform for gene regulation." Epigenetics: official journal of the DNA Methylation Society **6**(2): 147.

Blackledge, N. P., J. C. Zhou, et al. (2010). "CpG islands recruit a histone H3 lysine 36 demethylase." Molecular cell **38**(2): 179-190.

Blomen, V. A. and J. Boonstra (2011). "Stable transmission of reversible modifications: maintenance of epigenetic information through the cell cycle." Cellular and Molecular Life Sciences: 1-18.

Bock, C. and T. Lengauer (2008). "Computational epigenetics." Bioinformatics **24**(1): 1-10.

Bock, C., M. Paulsen, et al. (2006). "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure." PLoS genetics **2**(3): e26.

Bonasio, R., S. Tu, et al. (2010). "Molecular signals of epigenetic states." Science Signalling **330**(6004): 612.

Bork, S., S. Pfister, et al. (2010). "DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells." Aging cell **9**(1): 54-63.

Britten, R. J. and E. H. Davidson (1969). "Gene regulation for higher cells: a theory." Science **165**(891): 349-357.

Cain, C. E., R. Blekhman, et al. (2011). "Gene expression differences among primates are associated with changes in a histone epigenetic modification." Genetics **187**(4): 1225-34.

Celniker, S. E., L. A. Dillon, et al. (2009). "Unlocking the secrets of the genome." Nature **459**(7249): 927-30.

Celniker, S. E., L. A. L. Dillon, et al. (2009). "Unlocking the secrets of the genome." Nature **459**(7249): 927-930.

Chen, T., Y. Ueda, et al. (2003). "Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b." Molecular and cellular biology **23**(16): 5594-5605.

Chen, Z.-x. and A. D. Riggs (2011). "DNA methylation and demethylation in mammals." Journal of Biological Chemistry **286**(21): 18347-18353.

Clouaire, T. and I. Stancheva (2008). "Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin?" Cellular and Molecular Life Sciences **65**(10): 1509-1522.

Clouaire, T., S. Webb, et al. (2012). "Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells." Genes & Development **26**(15): 1714-1728.

Cocozza, S., M. M. Akhtar, et al. (2011). "CpG Islands Undermethylation in Human Genomic Regions under Selective Pressure." PloS one **6**(8): e23156.

Cohen, N. M., E. Kenigsberg, et al. (2011). "Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection." Cell **145**(5): 773.

Creyghton, M. P., A. W. Cheng, et al. (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state." Proceedings of the National Academy of Sciences **107**(50): 21931-21936.

Cropley, J. E., T. H. Y. Dang, et al. (2012). "The penetrance of an epigenetic trait in mice is progressively yet reversibly increased by selection and environment." Proceedings of the Royal Society B: Biological Sciences **279**(1737): 2347-2353.

Dawson, M. A. and T. Kouzarides (2012). "Cancer epigenetics: from mechanism to therapy." Cell **150**(1): 12-27.

Deaton, A. M. and A. Bird (2011). "CpG islands and the regulation of transcription." Genes & Development **25**(10): 1010.
Deaton, A. M., S. Webb, et al. (2011). "Cell type-specific DNA methylation at intragenic CpG islands in the immune system." Genome research **21**(7): 1074-1086.
Dou, Y., T. A. Milne, et al. (2006). "Regulation of MLL1 H3K4 methyltransferase activity by its core components." Nature structural & molecular biology **13**(8): 713-719.
Dunham, I., A. Kundaje, et al. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.
Duns, G., E. van den Berg, et al. (2010). "Histone methyltransferase gene SETD2 is a novel tumor suppressor gene in clear cell renal cell carcinoma." Cancer Research **70**(11): 4287-4291.
Duthie, S. J. (2011). "Symposium 1: Nutrition and epigenetics Epigenetic modifications and human pathologies: cancer and CVD." Proceedings of the Nutrition Society **70**(1): 47-56.
Eckhardt, F., J. Lewin, et al. (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature genetics **38**(12): 1378-1385.
Edmunds, J. W., L. C. Mahadevan, et al. (2007). "Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation." The EMBO journal **27**(2): 406-420.
Enard, W., A. Fassbender, et al. (2004). "Differences in DNA methylation patterns between humans and chimpanzees." Current Biology **14**(4): 148-149.
Ernst, J., P. Kheradpour, et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature **473**(7345): 43-9.
Feinberg, A. P. and B. Tycko (2004). "The history of cancer epigenetics." Nature Reviews Cancer **4**(2): 143-53.
Flanagan, J. M., V. Popendikyte, et al. (2006). "Intra-and interindividual epigenetic variation in human germ cells." The American Journal of Human Genetics **79**(1): 67-84.
Fouse, S. D., R. P. Nagarajan, et al. (2010). "Genome-scale DNA methylation analysis." Epigenomics **2**(1): 105-117.
Fraga, M. F., E. Ballestar, et al. (2005). "Epigenetic differences arise during the lifetime of monozygotic twins." Proceedings of the National Academy of Sciences of the United States of America **102**(30): 10604-10609.
Gama-Sosa, M. A., R. M. Midgett, et al. (1983). "Tissue-specific differences in DNA methylation in various mammals." Biochimica et biophysica acta **740**(2): 212.
Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." Journal of molecular biology **196**(2): 261.
Gibbs, J. R., M. P. Van Der Brug, et al. (2010). "Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain." PLoS genetics **6**(5): e1000952.
Gowher, H. and A. Jeltsch (2001). "Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpA sites." Journal of molecular biology **309**(5): 1201-1208.
Green, R. E., J. Krause, et al. (2010). "A draft sequence of the Neandertal genome." Science **328**(5979): 710-722.
Gregory, G. D., C. R. Vakoc, et al. (2007). "Mammalian ASH1L is a histone methyltransferase that occupies the transcribed region of active genes." Molecular and cellular biology **27**(24): 8466.
Guenther, M. G., S. S. Levine, et al. (2007). "A chromatin landmark and transcription initiation at most promoters in human cells." Cell **130**(1): 77-88.
Guttman, M., I. Amit, et al. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.
Haaf, T. (2006). "Methylation dynamics in the early mammalian embryo: implications of genome reprogramming defects for development." DNA Methylation: Development, Genetic Disease and Cancer: 13-22.
Han, L., P. D. W. Witmer, et al. (2007). "DNA methylation regulates MicroRNA expression." Cancer biology & therapy **6**(8): 1290-1294.
Handel, A. E., G. C. Ebers, et al. (2010). "Epigenetics: molecular mechanisms and implications for disease." Trends in molecular medicine **16**(1): 7-16.
Heijmans, B. T., D. Kremer, et al. (2007). "Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus." Human molecular genetics **16**(5): 547-554.
Holliday, R. and G. W. Grigg (1993). "DNA methylation and mutation." Mutat Res **285**: 61–67.
Illingworth, R. S. and A. P. Bird (2009). "CpG islands-"a rough guide"." FEBS letters **583**(11): 1713-1720.
Illingworth, R. S., U. Gruenewald-Schneider, et al. (2010). "Orphan CpG islands identify numerous conserved promoters in the mammalian genome." PLoS genetics **6**(9): e1001134.
Jablonka, E. and M. J. Lamb (1989). "The inheritance of acquired epigenetic variations." Journal of Theoretical Biology **139**(1): 69-83.

Jeong, S., G. Liang, et al. (2009). "Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA." Molecular and cellular biology **29**(19): 5366-5376.
Johnson, L. J. and P. J. Tricker (2010). "Epigenomic plasticity within populations: its evolutionary significance and potential." Heredity **105**(1): 113-121.
Jones, P. A. (2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond." Nature Reviews Genetics **13**(7): 484-92.
Jones, P. A. and G. Liang (2009). "Rethinking how DNA methylation patterns are maintained." Nature Reviews Genetics **10**(11): 805-811.
Kalkhoven, E. (2004). "CBP and p300: HATs for different occasions." Biochemical pharmacology **68**(6): 1145.
Kaneda, R., S. Takada, et al. (2009). "Genome-wide histone methylation profile for heart failure." Genes to Cells **14**(1): 69-77.
Ke, X. S., Y. Qu, et al. (2009). "Genome-wide profiling of histone h3 lysine 4 and lysine 27 trimethylation reveals an epigenetic signature in prostate carcinogenesis." PLoS One **4**(3): e4687.
Klose, R. J. and A. P. Bird (2006). "Genomic DNA methylation: the mark and its mediators." Trends in biochemical sciences **31**(2): 89-97.
Klose, R. J., K. Yamane, et al. (2006). "The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36." Nature **442**(7100): 312-316.
Kolasinska-Zwierz, P., T. Down, et al. (2009). "Differential chromatin marking of introns and expressed exons by H3K36me3." Nature Genetics **41**(3): 376-81.
Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.
Larsen, F., G. Gundersen, et al. (1992). "CpG islands as gene markers in the human genome." Genomics **13**(4): 1095-1107.
Law, J. A. and S. E. Jacobsen (2010). "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nature Reviews Genetics **11**(3): 204-20.
Lee, J. H. and D. G. Skalnik (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." Journal of Biological Chemistry **280**(50): 41725.
Li, B., M. Carey, et al. (2007). "The role of chromatin during transcription." Cell **128**(4): 707-719.
Li, B., J. Jackson, et al. (2009). "Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription." Journal of Biological Chemistry **284**(12): 7970-7976.
Liang, G., M. F. Chan, et al. (2002). "Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements." Molecular and cellular biology **22**(2): 480-491.
Lin, J. C., S. Jeong, et al. (2007). "Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island." Cancer cell **12**(5): 432-444.
Lister, R., M. Pelizzola, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." nature **462**(7271): 315-322.
Llamas, B., M. L. Holland, et al. (2012). "High-Resolution Analysis of Cytosine Methylation in Ancient DNA." PloS one **7**(1): e30226.
Lloret-Llinares, M., S. Pérez-Lluch, et al. (2012). "dKDM5/LID regulates H3K4me3 dynamics at the transcription-start site (TSS) of actively transcribed developmental genes." Nucleic acids research **279**(11): 1905-14.
Lopez-Serra, P. and M. Esteller (2012). "DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer." Oncogene **31**(13): 1609-1622.
Luger, K., A. W. Mader, et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution." Nature **389**(6648): 251-260.
Lund, A. H. and M. Van Lohuizen (2004). "Epigenetics and cancer." Genes & development **18**(19): 2315-2335.
Mancini, D. N., S. M. Singh, et al. (1999). "Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors." Oncogene **18**(28): 4108.
Maunakea, A. K., R. P. Nagarajan, et al. (2010). "Conserved role of intragenic DNA methylation in regulating alternative promoters." Nature **466**(7303): 253-257.
Medvedeva, Y. A., M. V. Fridman, et al. (2010). "Intergenic, gene terminal, and intragenic CpG islands in the human genome." BMC genomics **11**: 48.
Meehan, R. R., J. D. Lewis, et al. (1989). "Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs." Cell **58**(3): 499-507.
Meissner, A., T. S. Mikkelsen, et al. (2008). "Genome-scale DNA methylation maps of pluripotent and differentiated cells." Nature **454**(7205): 766-770.
Mikkelsen, T. S., J. Hanna, et al. (2008). "Dissecting direct reprogramming through integrative genomic analysis." Nature **454**(7200): 49-55.
Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Miller, W., D. I. Drautz, et al. (2008). "Sequencing the nuclear genome of the extinct woolly mammoth." Nature **456**(7220): 387-390.
Mohn, F., M. Weber, et al. (2008). "Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors." Molecular cell **30**(6): 755-766.
Murrell, A., S. Heeson, et al. (2004). "An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype." Human molecular genetics **13**(2): 247-255.
Nan, X., H.-H. Ng, et al. (1998). "Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex." Nature **393**(6683): 386-389.
Newbold, R. F. and K. Mokbel (2010). "Evidence for a tumour suppressor function of SETD2 in human breast cancer: a new hypothesis." Anticancer research **30**(9): 3309-3311.
Nielsen, R. (2005). "Molecular signatures of natural selection." Annual Review of Genetics **39**: 197-218.
Okano, M., D. W. Bell, et al. (1999). "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development." Cell **99**(3): 247-257.
Okano, M., S. Xie, et al. (1998). "Cloning and characterization of a family of novel mammalian DNA." Nat. Genet **19**(3): 219-220.
Okitsu, C. Y. and C. L. Hsieh (2007). "DNA methylation dictates histone H3K4 methylation." Molecular and cellular biology **27**(7): 2746.
Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature Reviews Genetics **10**(10): 669-680.
Pasini, D., M. Malatesta, et al. (2010). "Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes." Nucleic acids research **38**(15): 4958-4969.
Payer, B. and J. T. Lee (2008). "X chromosome dosage compensation: how mammals keep the balance." Annual review of genetics **42**: 733-772.
Pelizzola, M. and J. R. Ecker (2011). "The DNA methylome." FEBS letters **585**(1994–2000).
Petronis, A. (2010). "Epigenetics as a unifying principle in the aetiology of complex traits and diseases." Nature **465**(7299): 721-727.
Pickrell, J. K., G. Coop, et al. (2009). "Signals of recent positive selection in a worldwide sample of human populations." Genome Research **19**(5): 826-837.
Pollard, K. S., M. J. Hubisz, et al. (2010). "Detection of nonneutral substitution rates on mammalian phylogenies." Genome research **20**(1): 110-121.
R_Development_Core_Team (2009). "R: a language and environment for statistical computing." Vienna, Austria

Ramser, J., M. E. Ahearn, et al. (2008). "Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy." The American Journal of Human Genetics **82**(1): 188-193.
Reich, D., R. E. Green, et al. (2010). "Genetic history of an archaic hominin group from Denisova Cave in Siberia." Nature **468**(7327): 1053-1060.
Renda, M., I. Baglivo, et al. (2007). "Critical DNA Binding Interactions of the Insulator Protein CTCF." Journal of Biological Chemistry **282**(46): 33336-33345.
Reynolds, N., M. Salmon-Divon, et al. (2011). "NuRD-mediated deacetylation of H3K27 facilitates recruitment of Polycomb Repressive Complex 2 to direct gene repression." The EMBO journal **31**(3): 593-605.
Richards, E. J. (2006). "Inherited epigenetic variation-revisiting soft inheritance." Nature Reviews Genetics **7**(5): 395-401.
Richards, E. J. (2008). "Population epigenetics." Current opinion in genetics & development **18**(2): 221-226.
Riggs, A. D. and Z. Xiong (2004). "Methylation and epigenetic fidelity." Proceedings of the National Academy of Sciences **101**(1): 4-5.
Roh, T.-y., G. Wei, et al. (2007). "Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns." Genome research **17**(1): 74-81.
Ruthenburg, A. J., C. D. Allis, et al. (2007). "Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark." Molecular cell **25**(1): 15-30.
Sabeti, P. C., D. E. Reich, et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-837.
Sabeti, P. C., P. Varilly, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-918.
Saferali, A., E. Grundberg, et al. (2010). "Cell culture-induced aberrant methylation of the imprinted IG DMR in human lymphoblastoid cell lines." Epigenetics **5**(1): 144-154.
Sanborn, J. Z., S. C. Benz, et al. (2011). "The UCSC cancer genomics browser: update 2011." Nucleic acids research **39**(suppl 1): D951-D959.

Sandgren, J., R. Andersson, et al. (2010). "Integrative epigenomic and genomic analysis of malignant pheochromocytoma." Experimental & molecular medicine **42**(7): 484.
Saxonov, S., P. Berg, et al. (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." Proceedings of the National Academy of Sciences of the United States of America **103**(5): 1412-1417.
Schlesinger, Y., R. Straussman, et al. (2006). "Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer." Nature genetics **39**(2): 232-236.
Sharif, J., M. Muto, et al. (2007). "The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA." Nature **450**(7171): 908-912.
Shi, Y., F. Lan, et al. (2004). "Histone demethylation mediated by the nuclear amine oxidase homolog LSD1." Cell **119**(7): 941-953.
Shulha, H. P., J. L. Crisci, et al. (2012). "Human-Specific Histone Methylation Signatures at Transcription Start Sites in Prefrontal Neurons." PLoS Biology **10**(11): e1001427.
Siepel, A., G. Bejerano, et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Research **15**(8): 1034-50.
Smilinich, N. J., C. D. Day, et al. (1999). "A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwithâ€"Wiedemann syndrome." Proceedings of the National Academy of Sciences **96**(14): 8064-8069.
Soojin, V. Y. and M. A. D. Goodisman (2009). "Computational approaches for understanding the evolution of DNA methylation in animals." Epigenetics **4**(8): 551-556.
Stein, R., A. Razin, et al. (1982). "In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells." Proceedings of the National Academy of Sciences **79**(11): 3418-3422.
Strahl, B. D., P. A. Grant, et al. (2002). "Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression." Molecular and cellular biology **22**(5): 1298-1306.
Su, J., Y. Zhang, et al. (2010). "CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes." Nucleic acids research **38**(1): e6-e6.
Takai, D. and P. A. Jones (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proceedings of the national academy of sciences **99**(6): 3740-3745.
Tammen, S. A., S. Friso, et al. (2012). "Epigenetics: the link between nature and nurture." Molecular Aspects of Medicine **in press**.
Tazi, J. and A. Bird (1990). "Alternative chromatin structure at CpG islands." Cell **60**(6): 909.
Thomson, J. P., P. J. Skene, et al. (2010). "CpG islands influence chromatin structure via the CpG-binding protein Cfp1." Nature **464**(7291): 1082-1086.
Tie, F., R. Banerjee, et al. (2009). "CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing." Development **136**(18): 3131-3141.
Tomso, D. J. and D. A. Bell (2003). "Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands." Journal of molecular biology **327**(2): 303-308.
Tsukada, Y., J. Fang, et al. (2006). "Histone demethylation by a family of JmjC domain-containing proteins." Nature **439**(7078): 811-6.
Vaughn, M. W., M. Tanurdžić, et al. (2007). "Epigenetic natural variation in Arabidopsis thaliana." PLoS biology **5**(7): e174.
Vernot, B., A. B. Stergachis, et al. (2012). "Personal and population genomics of human regulatory variation." Genome Res **22**(9): 1689-97.
Viré, E., C. Brenner, et al. (2005). "The Polycomb group protein EZH2 directly controls DNA methylation." Nature **439**(7078): 871-874.
Voight, B. F., S. Kudaravalli, et al. (2006). "A map of recent positive selection in the human genome." PLoS Biology **4**(3): e72.
Voo, K. S., D. L. Carlone, et al. (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." Molecular and Cellular Biology **20**(6): 2108.
Waddington, C. H. (1942). "The epigenotype." Endeavour **1**: 18–20.
Wagner, E. J. and P. B. Carpenter (2012). "Understanding the language of Lys36 methylation at histone H3." Nature Reviews Molecular Cell Biology **13**(2): 115-126.
Wang, F., C. B. Marshall, et al. (2013). "Transcriptional/epigenetic regulator CBP/p300 in tumorigenesis: structural and functional versatility in target recognition." Cellular and Molecular Life Sciences **In press**: 1-20.
Wang, G., Y. Wang, et al. (2010). "RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation." PloS one **5**(11): e13798.

Wang, Z., C. Zang, et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nature genetics **40**(7): 897-903.
Wilson, M. D., N. L. Barbosa-Morais, et al. (2008). "Species-specific transcription in mice carrying human chromosome 21." Science **322**(5900): 434-438.
Woo, Y. H. and W. H. Li (2012). "Evolutionary conservation of histone modifications in mammals." Molecular Biology and Evolution **29**(7): 1757-67.
Zhang, D., L. Cheng, et al. (2010). "Genetic control of individual differences in gene-specific methylation in human brain." The American Journal of Human Genetics **86**(3): 411-419.
Zhang, X., S. Shiu, et al. (2008). "Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays." PLoS genetics **4**(3): e1000032.
Zhu, J., F. He, et al. (2008). "On the nature of human housekeeping genes." Trends in Genetics **24**(10): 481-484.

Supplementary informations

Supplementary table 2.1: Complete list of the cell used in this study, with their characteristics

Cell ID	Category	Tissue	Cell Type
Hek293	cancer	embryonic kidney	various cell types (mostly neuronal)
MCF-7	cancer	pleural effusion	luminal epithelial
Hepg2	cancer	Epithelial Cells	endoderm
Cmk	cancer	peripheral blood	megakaryocytes
NB4	cancer	stromal fibroblasts	promyelocytes
NT2-D1	cancer	testis	Epithelial Cells
Gm19239	EBV	Blood	B-Lymphocyte
Gm19240	EBV	Blood	B-Lymphocyte
Ag04449	normal	Skin	Fibroblast
Ag04450	normal	Lung	Fibroblast
Ag09309	normal	Skin	Fibroblast
Ag09319	normal	Gingival	Fibroblast
Ag10803	normal	Skin	Fibroblast
FibrobI	normal	Skin	Fibroblast
HAEpiC	normal	Epithelial Cells	Endothelial Cells
HCF	normal	heart	Fibroblast
HCM	normal	heart	Myocytes
HEEpiC	normal	Epithelial Cells	Epithelial Cells
HIPEpiC	normal	Iris Pigment Cells	Epithelial Cells
HMEC	normal	Epithelial Cells	Epithelial Cells
HNPCEpiC	normal	basal membranes(NPCEC)	ciliary epithelium
HRCEpiC	normal	renal epithelial cells	epithelial cells
HSMMtube	normal	Skeletal Muscle	myoblasts and myotubes
NHBE	normal	Bronchial/Tracheal Epithelial cells	epithelial cells
Skmc	normal	Muscle Cells	Muscle Cells

Supplementary table 2.2: Lists, for each cell type, the number of CpG analyzed, the number of CpGs inside CGIs, the number of CGIs for which we were able to estimate methylation, the number of CpG analyzed per CGI and the mean value of CGI methylation.

Cell line	Total no. of CpG analyzed	No. of CpGs inside the CGIs	No. of CGIs with estimated methylation	No. of CpG analyzed/CGI	Mean value of CGI methylation
Hek293	556343	410965	16821	24.43	21.35
MCF-7	691911	477299	18199	26.23	32.77
Hepg2	632281	445920	18191	24.51	27.02
Cmk	668036	469564	17843	26.32	32.93
NB4	671651	482994	17845	27.07	31.47
NT2-D1	546153	414540	16471	25.17	15.95
Gm19239	520138	374026	16858	22.19	17.47
Gm19240	683988	476468	18577	25.65	20.39
Ag04449	601041	467048	17108	27.30	9.96
Ag04450	630183	447284	18043	24.79	13.66
Ag09309	691076	493539	18113	27.25	16.79
Ag09319	545847	396283	17150	23.11	14.44
Ag10803	774389	535976	19161	27.97	15.66
Fibrobl	632283	451557	17608	25.64	17.17
HAEpiC	681081	495882	18034	27.50	13.75
HCF	512497	378981	16137	23.49	12.6
HCM	760148	548934	18646	29.44	13.29
HEEpiC	624594	449729	18644	24.12	13.06
HIPEpiC	628948	459947	17722	25.95	13.22
HMEC	635871	460372	17825	25.83	16
HNPCEpiC	718881	530483	18431	28.78	12.8
HRCEpiC	505006	374743	16807	22.30	11.72
HSMMtube	704265	493549	14869	33.19	19.48
NHBE	680749	491581	18156	27.08	14.18
Skmc	674758	477801	16471	29.01	15.95

Supplementary table 2.3: Lists, for each HIR identified, the chromosome, the start position, the end position, the total length and the human population in which it has been detected. Genomic coordinates refer to assembly GRCh37/hg19.

Chromosome	Start	End	Length	Population
chr1	202566430	202843126	276696	Bantu
chr1	113828836	114587765	758929	Bantu
chr1	175113118	175336529	223411	Bantu
chr1	160838638	161232771	394133	Bantu
chr1	37023791	37307280	283489	Bantu
chr1	79217345	79519426	302081	Bantu
chr1	88983612	89384716	401104	Bantu
chr1	35322503	36756490	1433987	Middle East
chr1	229851411	230250181	398770	Middle East
chr1	155059851	156078248	1018397	Middle East
chr1	74560411	75027257	466846	Middle East
chr1	154795389	156148164	1352775	East Asia
chr1	65661665	66225796	564131	East Asia
chr1	8320045	9000881	680836	East Asia
chr1	172498216	173266577	768361	East Asia
chr1	234701860	234906828	204968	East Asia
chr1	30348925	30562065	213140	East Asia
chr1	75098453	76399205	1300752	East Asia
chr1	23184206	23834485	650279	East Asia
chr1	64163143	64490892	327749	East Asia
chr1	76463072	76729910	266838	East Asia
chr1	81474889	81908616	433727	East Asia
chr1	92464663	93523082	1058419	East Asia
chr1	234637347	234853058	215711	South Asia
chr1	186205867	186700958	495091	South Asia
chr1	219644224	220082149	437925	South Asia
chr1	236708968	236993127	284159	South Asia
chr1	248004775	248344597	339822	Europe
chr1	35321242	36833079	1511837	Europe
chr1	161950451	162303458	353007	Europe
chr1	186164899	186719569	554670	Europe
chr1	30638016	30791524	153508	Europe
chr1	184905261	185437237	531976	Europe
chr1	220535456	220934076	398620	Europe
chr1	11419868	11736599	316731	Europe
chr1	89415576	90132068	716492	America
chr1	161263380	162008078	744698	America
chr2	155649338	156309902	660564	Bantu
chr2	72353138	73170729	817591	Bantu
chr2	212010493	212199993	189500	Bantu
chr2	238592234	238789194	196960	Bantu
chr2	209141066	209434192	293126	Bantu
chr2	33174806	33338735	163929	Bantu

chr2	46622873	46748065	125192	Bantu
chr2	140962703	141212782	250079	Bantu
chr2	200751772	201256573	504801	Bantu
chr2	72239692	73220137	980445	Middle East
chr2	1752229	2163075	410846	Middle East
chr2	178148204	178579812	431608	Middle East
chr2	195377437	195968049	590612	Middle East
chr2	14756349	15076200	319851	Middle East
chr2	74463566	75012326	548760	Middle East
chr2	235709572	235971970	262398	Middle East
chr2	21587225	22197491	610266	Middle East
chr2	117838706	118274001	435295	Middle East
chr2	157923137	158915734	992597	Middle East
chr2	108546122	109825515	1279393	East Asia
chr2	16798475	18056446	1257971	East Asia
chr2	212788895	213695701	906806	East Asia
chr2	84536960	85288571	751611	East Asia
chr2	125232609	127372727	2140118	East Asia
chr2	154510972	155053139	542167	East Asia
chr2	8794240	9870630	1076390	East Asia
chr2	177293076	179236047	1942971	East Asia
chr2	223503370	223856275	352905	East Asia
chr2	43638712	44129773	491061	East Asia
chr2	177914806	178612845	698039	South Asia
chr2	231638974	231869040	230066	South Asia
chr2	195552753	197098637	1545884	South Asia
chr2	223549297	224218792	669495	South Asia
chr2	157993888	158922574	928686	South Asia
chr2	82644948	83438063	793115	South Asia
chr2	73590806	74119507	528701	South Asia
chr2	178158043	178612269	454226	Europe
chr2	167525345	168461160	935815	Europe
chr2	197255486	198056040	800554	Europe
chr2	195198744	196465135	1266391	Europe
chr2	74461634	75008283	546649	Europe
chr2	196622252	197143630	521378	Europe
chr2	158095057	158890761	795704	Europe
chr2	121152814	121332190	179376	Europe
chr2	141536876	141803601	266725	Europe
chr2	17432556	18185589	753033	America
chr2	169610232	169870854	260622	America
chr2	36045002	36399317	354315	America
chr2	39776534	40283458	506924	America
chr2	102288186	103208609	920423	America
chr3	4445436	4593187	147751	Bantu
chr3	29582412	29735362	152950	Bantu
chr3	56467691	56858493	390802	Bantu

chr3	61066987	61267415	200428	Bantu
chr3	76685884	76869830	183946	Bantu
chr3	30263217	30416461	153244	Bantu
chr3	63029516	63168593	139077	Bantu
chr3	105006914	105670159	663245	Bantu
chr3	25495637	26448193	952556	Middle East
chr3	5547942	5777221	229279	Middle East
chr3	148640368	148947187	306819	Middle East
chr3	175042776	175530096	487320	East Asia
chr3	104618616	104903304	284688	East Asia
chr3	44091635	45170504	1078869	East Asia
chr3	195606770	196351214	744444	East Asia
chr3	25666156	26450380	784224	East Asia
chr3	26976847	27638499	661652	East Asia
chr3	112911525	113233187	321662	East Asia
chr3	166170696	167890717	1720021	East Asia
chr3	25489318	26447274	957956	South Asia
chr3	140485038	141050804	565766	South Asia
chr3	129699722	130468624	768902	South Asia
chr3	165267688	166012937	745249	South Asia
chr3	12481375	12857492	376117	South Asia
chr3	66744399	66996178	251779	South Asia
chr3	101816344	102457117	640773	South Asia
chr3	182957573	183177706	220133	South Asia
chr3	72533768	72667610	133842	Europe
chr3	129658706	130477565	818859	Europe
chr3	190295730	190629384	333654	Europe
chr3	156259485	156821807	562322	America
chr3	100561439	100892053	330614	America
chr3	20984501	21456169	471668	America
chr3	63663692	64099568	435876	America
chr4	118796265	119853637	1057372	Bantu
chr4	41798793	42165798	367005	Bantu
chr4	129779076	130144490	365414	Bantu
chr4	33653757	35047016	1393259	Bantu
chr4	65507693	65895584	387891	Bantu
chr4	429720	871674	441954	Bantu
chr4	86088554	86358620	270066	Bantu
chr4	111540391	111765494	225103	Bantu
chr4	33756410	34855431	1099021	Middle East
chr4	169771612	170787253	1015641	Middle East
chr4	148032003	148987971	955968	Middle East
chr4	14442433	14861269	418836	Middle East
chr4	29775276	30429541	654265	Middle East
chr4	41322398	42206573	884175	Middle East
chr4	172311051	172742576	431525	Middle East
chr4	104890866	105457179	566313	East Asia

chr4	143493213	144542212	1048999	East Asia
chr4	152792791	153097095	304304	East Asia
chr4	157656264	160680412	3024148	East Asia
chr4	41322398	42244444	922046	East Asia
chr4	169669891	170762016	1092125	East Asia
chr4	41407668	42241035	833367	South Asia
chr4	5171986	5426326	254340	South Asia
chr4	173213263	173863771	650508	South Asia
chr4	179935508	180172955	237447	South Asia
chr4	32672332	34855431	2183099	Europe
chr4	5218136	5425361	207225	Europe
chr4	41333429	42227496	894067	Europe
chr4	171743216	172783333	1040117	Europe
chr4	148025559	148491479	465920	Europe
chr4	14438997	14862872	423875	Europe
chr4	123228113	124245085	1016972	Europe
chr4	42331579	42627835	296256	America
chr4	126738833	127236804	497971	America
chr4	128495654	129509264	1013610	America
chr5	108964907	109272681	307774	Bantu
chr5	160174288	160380400	206112	Bantu
chr5	66794632	67121200	326568	Bantu
chr5	87125578	88021526	895948	Bantu
chr5	169485227	169625918	140691	Bantu
chr5	25495542	25691581	196039	Bantu
chr5	142031751	142520585	488834	Middle East
chr5	24497516	24847070	349554	Middle East
chr5	11366401	11893756	527355	Middle East
chr5	19493684	20429331	935647	Middle East
chr5	109401940	110455898	1053958	Middle East
chr5	30028447	30913502	885055	Middle East
chr5	92219182	92675984	456802	Middle East
chr5	90466296	90782146	315850	Middle East
chr5	37913527	38135717	222190	East Asia
chr5	97821452	98569708	748256	East Asia
chr5	107994013	108579888	585875	East Asia
chr5	113425033	113847911	422878	East Asia
chr5	172845597	173103336	257739	East Asia
chr5	92595749	93637332	1041583	East Asia
chr5	54601452	55204186	602734	South Asia
chr5	109636606	110500986	864380	South Asia
chr5	37909151	38166348	257197	South Asia
chr5	158944085	159230739	286654	South Asia
chr5	120960853	121265924	305071	South Asia
chr5	142031751	142238118	206367	South Asia
chr5	165886820	166090938	204118	South Asia
chr5	108031617	108581017	549400	South Asia

chr5	109475723	110399880	924157	Europe
chr5	21541191	22083472	542281	Europe
chr5	37915147	38165207	250060	Europe
chr5	24486805	24851079	364274	Europe
chr5	142020380	142529959	509579	Europe
chr5	80687791	81327041	639250	America
chr5	140553664	141326961	773297	America
chr5	164283543	164865385	581842	America
chr5	59556461	60224913	668452	America
chr5	153277759	153988723	710964	America
chr6	74981125	75538209	557084	Bantu
chr6	76857008	77299932	442924	Bantu
chr6	130503306	130789540	286234	Bantu
chr6	72662228	73018440	356212	Bantu
chr6	168410476	168546851	136375	Bantu
chr6	40420901	40616865	195964	Bantu
chr6	70802259	70990469	188210	Bantu
chr6	132919948	133067732	147784	Bantu
chr6	97538657	98044631	505974	Bantu
chr6	73650662	73866941	216279	Middle East
chr6	107420270	108089687	669417	Middle East
chr6	1045539	1149482	103943	Middle East
chr6	4744063	4984230	240167	Middle East
chr6	14688469	15016186	327717	Middle East
chr6	16109163	16362387	253224	Middle East
chr6	123902541	124158251	255710	Middle East
chr6	25352637	27356924	2004287	East Asia
chr6	105369510	106017975	648465	East Asia
chr6	37278933	37585912	306979	East Asia
chr6	73650662	74064395	413733	South Asia
chr6	105105089	105977159	872070	South Asia
chr6	213983	412740	198757	South Asia
chr6	104765922	106006229	1240307	Europe
chr6	73652443	73893196	240753	Europe
chr6	167169223	167521623	352400	Europe
chr6	121812866	123188087	1375221	America
chr6	46859043	47196209	337166	America
chr6	46283922	46745998	462076	America
chr6	152774033	153103098	329065	America
chr6	158382074	158865656	483582	America
chr6	33889404	36757265	2867861	America
chr6	49902734	50629153	726419	America
chr6	72806978	73161392	354414	America
chr6	150748277	150945214	196937	America
chr7	40485949	40946512	460563	Bantu
chr7	2739017	2966137	227120	Bantu
chr7	141214012	141710898	496886	Bantu

chr7	28753387	28926167	172780	Bantu
chr7	91998495	92480883	482388	Bantu
chr7	17912336	18094286	181950	Bantu
chr7	115654075	115851980	197905	Bantu
chr7	123656664	124152389	495725	Bantu
chr7	156416220	156754980	338760	Bantu
chr7	118610284	120731948	2121664	Middle East
chr7	124182414	124827235	644821	Middle East
chr7	33555119	33839113	283994	Middle East
chr7	97626161	98123879	497718	Middle East
chr7	98772131	99447045	674914	Middle East
chr7	3726122	4333609	607487	East Asia
chr7	49640912	50313351	672439	East Asia
chr7	28693593	28934502	240909	East Asia
chr7	29833587	30217699	384112	East Asia
chr7	64807261	66384314	1577053	East Asia
chr7	111811004	112304166	493162	East Asia
chr7	126457779	127852064	1394285	East Asia
chr7	138518783	138794878	276095	East Asia
chr7	117496319	120683750	3187431	South Asia
chr7	36969211	37341476	372265	South Asia
chr7	148636850	149017787	380937	South Asia
chr7	36818523	37260711	442188	Europe
chr7	98759117	99525240	766123	Europe
chr7	119301786	120683750	1381964	Europe
chr7	102486254	103231289	745035	Europe
chr7	107948721	108262713	313992	America
chr7	99300758	100968362	1667604	America
chr7	122627567	123037886	410319	America
chr7	92623141	93013573	390432	America
chr8	21164751	21311469	146718	Bantu
chr8	113645627	114021483	375856	Bantu
chr8	9410495	9681256	270761	Bantu
chr8	68653516	68838275	184759	Bantu
chr8	99766557	100855842	1089285	Bantu
chr8	5225481	5409370	183889	Bantu
chr8	9087278	9243427	156149	Bantu
chr8	139901178	140127522	226344	Middle East
chr8	18514100	18673191	159091	Middle East
chr8	59793692	59962803	169111	Middle East
chr8	36629798	37046811	417013	Middle East
chr8	129503361	130268559	765198	East Asia
chr8	134804170	135328581	524411	East Asia
chr8	21616454	21988514	372060	South Asia
chr8	57826145	58251654	425509	South Asia
chr8	66555818	67019987	464169	South Asia
chr8	42045655	43215239	1169584	South Asia

chr8	139926589	140131393	204804	Europe
chr8	29692897	30218244	525347	Europe
chr8	18511097	18673191	162094	Europe
chr8	32507149	33806909	1299760	America
chr8	99383408	100997713	1614305	America
chr8	95303456	95847149	543693	America
chr8	707884	1099985	392101	America
chr9	23785967	24654574	868607	Bantu
chr9	115404746	115678032	273286	Bantu
chr9	114868003	115347087	479084	Bantu
chr9	100507159	100871716	364557	Bantu
chr9	3862693	3961491	98798	Bantu
chr9	95374368	95855099	480731	Bantu
chr9	96447721	96681387	233666	Bantu
chr9	111693132	111961388	268256	Bantu
chr9	76765949	77038958	273009	Bantu
chr9	139647074	140376667	729593	Bantu
chr9	8787733	8868683	80950	Bantu
chr9	1707518	2165757	458239	Middle East
chr9	107870014	108376435	506421	Middle East
chr9	126154354	126768047	613693	Middle East
chr9	12935290	13426560	491270	Middle East
chr9	111537522	111961388	423866	Middle East
chr9	136658987	136807769	148782	Middle East
chr9	22986049	23425709	439660	East Asia
chr9	126153636	126783846	630210	East Asia
chr9	26506859	27148474	641615	East Asia
chr9	106554028	106960541	406513	East Asia
chr9	13780058	14014967	234909	East Asia
chr9	111114252	111973019	858767	East Asia
chr9	130755693	131580743	825050	South Asia
chr9	1784090	2160104	376014	South Asia
chr9	126154354	126769290	614936	South Asia
chr9	107878896	108244575	365679	South Asia
chr9	107869514	108315068	445554	Europe
chr9	3161347	3579890	418543	Europe
chr9	93928416	94448776	520360	Europe
chr9	126154354	126749630	595276	Europe
chr9	130908044	131595406	687362	Europe
chr9	111639682	111957172	317490	Europe
chr9	9729134	9893070	163936	Europe
chr9	16046983	16252409	205426	Europe
chr9	18944389	19144190	199801	Europe
chr9	12925921	13755191	829270	Europe
chr9	93946379	94498549	552170	America
chr9	113300206	113527564	227358	America
chr10	85960149	86356554	396405	Bantu

chr10	90042154	90436200	394046	Bantu
chr10	69021184	69430997	409813	Bantu
chr10	100339741	101069351	729610	Bantu
chr10	83577099	84424077	846978	Middle East
chr10	118047298	118351904	304606	Middle East
chr10	131166636	131498390	331754	Middle East
chr10	21783634	22926992	1143358	Middle East
chr10	53139938	53415187	275249	East Asia
chr10	3877968	4469291	591323	East Asia
chr10	107027076	107533730	506654	East Asia
chr10	109428376	110348375	919999	East Asia
chr10	73656982	74450318	793336	East Asia
chr10	93348014	95233855	1885841	East Asia
chr10	83583540	84386092	802552	South Asia
chr10	59386322	60255968	869646	South Asia
chr10	131160866	131494256	333390	South Asia
chr10	100225657	101185829	960172	South Asia
chr10	114912534	115198442	285908	Europe
chr10	118052051	118381944	329893	Europe
chr10	83835639	84386972	551333	Europe
chr10	6768836	6930654	161818	Europe
chr10	112570243	112918572	348329	Europe
chr10	109950470	110410029	459559	America
chr10	116565387	117735418	1170031	America
chr11	110336182	110796084	459902	Bantu
chr11	112727378	113103995	376617	Bantu
chr11	9791579	10432369	640790	Bantu
chr11	6043277	6301176	257899	Bantu
chr11	129392269	129568491	176222	Bantu
chr11	83638942	84024745	385803	Middle East
chr11	129832270	130264277	432007	Middle East
chr11	20702034	20979130	277096	Middle East
chr11	12157657	12285089	127432	Middle East
chr11	37977241	38407175	429934	Middle East
chr11	24839058	25755820	916762	East Asia
chr11	66688100	67438194	750094	East Asia
chr11	80930312	81515166	584854	East Asia
chr11	112812563	113056322	243759	East Asia
chr11	87342854	87674034	331180	South Asia
chr11	39439328	40043059	603731	South Asia
chr11	129832270	130264277	432007	South Asia
chr11	92108243	92623348	515105	South Asia
chr11	105719522	106288553	569031	South Asia
chr11	66240882	68468669	2227787	South Asia
chr11	87187306	87678365	491059	Europe
chr11	66687863	67414491	726628	Europe
chr11	86403375	86586679	183304	Europe

chr11	69231796	69579069	347273	America
chr11	6577577	6885406	307829	America
chr11	70966997	72935824	1968827	America
chr11	42864284	43656534	792250	America
chr12	28118847	28290777	171930	Bantu
chr12	5374517	5555039	180522	Bantu
chr12	81966423	82312036	345613	Bantu
chr12	113458677	113935480	476803	Bantu
chr12	126017498	126322479	304981	Middle East
chr12	99265706	99870626	604920	Middle East
chr12	10069302	10492981	423679	Middle East
chr12	127043455	127329166	285711	Middle East
chr12	121196891	121669181	472290	Middle East
chr12	888428	1644884	756456	East Asia
chr12	100097346	100868682	771336	East Asia
chr12	24451627	24925238	473611	East Asia
chr12	80462262	81032937	570675	South Asia
chr12	99332599	99870626	538027	South Asia
chr12	850613	1510504	659891	South Asia
chr12	95184157	95735182	551025	South Asia
chr12	59128693	59494105	365412	South Asia
chr12	65827996	66140732	312736	South Asia
chr12	2609562	2919700	310138	Europe
chr12	11460063	11714653	254590	Europe
chr12	10046052	10424626	378574	Europe
chr12	126039433	126290696	251263	Europe
chr12	99638276	99866158	227882	Europe
chr12	102389496	102838514	449018	Europe
chr12	47618366	48257765	639399	America
chr12	3723673	4042533	318860	America
chr13	30462712	30621645	158933	Bantu
chr13	46755852	47132369	376517	Bantu
chr13	104067224	104357021	289797	Middle East
chr13	33191371	33553782	362411	Middle East
chr13	74840712	75105119	264407	Middle East
chr13	104989752	105265994	276242	East Asia
chr13	34783249	35085471	302222	East Asia
chr13	103992269	104361190	368921	South Asia
chr13	67731496	68459217	727721	South Asia
chr13	92250545	92847826	597281	South Asia
chr13	45401128	45900266	499138	South Asia
chr13	103989946	104354010	364064	Europe
chr13	88900080	89934018	1033938	America
chr13	38429073	38803973	374900	America
chr14	60185336	61572948	1387612	Bantu
chr14	98497054	98776337	279283	Bantu
chr14	32007373	32391208	383835	Bantu

chr14	43480483	44259508	779025	Bantu
chr14	48570466	49066430	495964	Bantu
chr14	33942839	34054942	112103	Bantu
chr14	36524098	36809925	285827	Bantu
chr14	72658681	72892393	233712	Bantu
chr14	61662167	62159911	497744	Middle East
chr14	100852820	101247031	394211	Middle East
chr14	55218435	55888532	670097	Middle East
chr14	87470029	87984703	514674	East Asia
chr14	64873599	65660501	786902	South Asia
chr14	62536948	64673866	2136918	South Asia
chr14	102238630	102900767	662137	South Asia
chr14	24406306	24903963	497657	Europe
chr14	100852820	101240529	387709	Europe
chr14	100781877	101200644	418767	America
chr15	61994586	62397233	402647	Bantu
chr15	50944589	51394410	449821	Bantu
chr15	56337527	56659827	322300	Bantu
chr15	95284161	95544856	260695	Bantu
chr15	67049507	67234339	184832	Middle East
chr15	69178312	69991416	813104	Middle East
chr15	48132379	49038469	906090	Middle East
chr15	93984201	94225522	241321	Middle East
chr15	93293836	93648497	354661	East Asia
chr15	67029307	67227210	197903	South Asia
chr15	69424143	69991613	567470	South Asia
chr15	91716130	91864817	148687	South Asia
chr15	48224971	48812019	587048	Europe
chr15	42744094	43796907	1052813	Europe
chr15	72060049	73086730	1026681	Europe
chr15	58210398	58601803	391405	America
chr15	59655578	60363167	707589	America
chr15	72299481	73109628	810147	America
chr15	24267426	24999290	731864	America
chr16	81365974	81595185	229211	Bantu
chr16	22915924	23274063	358139	Bantu
chr16	12655483	12704564	49081	Bantu
chr16	25932733	26101005	168272	Bantu
chr16	82987188	83162677	175489	Middle East
chr16	1522670	2010137	487467	Middle East
chr16	24609208	24888453	279245	Middle East
chr16	65519567	65914253	394686	East Asia
chr16	79679369	79981629	302260	East Asia
chr16	17131750	17761858	630108	East Asia
chr16	75492242	76018395	526153	East Asia
chr16	79741569	79991174	249605	Europe
chr16	82228076	82456634	228558	Europe

chr16	10928112	11693535	765423	America
chr16	13594921	13853466	258545	America
chr16	82120192	82296370	176178	America
chr16	11919959	12314450	394491	America
chr16	70723925	73016767	2292842	America
chr17	3423578	3769406	345828	Bantu
chr17	45278790	45908404	629614	Bantu
chr17	49929451	50258110	328659	Bantu
chr17	19613601	20699544	1085943	Bantu
chr17	57637803	59345378	1707575	Middle East
chr17	55274630	55522449	247819	Middle East
chr17	63123532	63541496	417964	Middle East
chr17	74574103	74916503	342400	Middle East
chr17	53573902	54158281	584379	Middle East
chr17	76890864	77110043	219179	Middle East
chr17	48367272	48860431	493159	East Asia
chr17	27322441	29161357	1838916	East Asia
chr17	57832391	59423419	1591028	East Asia
chr17	63155764	63505321	349557	South Asia
chr17	53563973	54094316	530343	South Asia
chr17	63109056	63515782	406726	Europe
chr17	53519844	54106536	586692	Europe
chr17	74563175	74938080	374905	Europe
chr17	9390389	9708955	318566	Europe
chr17	76885117	77104322	219205	Europe
chr17	38775805	39650803	874998	America
chr17	35020291	35826181	805890	America
chr18	42619149	42883958	264809	Bantu
chr18	43959703	44146835	187132	Bantu
chr18	37544937	37836985	292048	Bantu
chr18	7384596	7706346	321750	Middle East
chr18	66570423	66881411	310988	Middle East
chr18	31042855	31383570	340715	East Asia
chr18	66587607	66871598	283991	South Asia
chr18	32009672	32375776	366104	South Asia
chr18	57598173	57767634	169461	South Asia
chr18	46076096	46240710	164614	South Asia
chr18	7147079	7732916	585837	Europe
chr18	66574279	66887310	313031	Europe
chr19	51907788	52127052	219264	Bantu
chr19	38602214	39002139	399925	Bantu
chr19	42207625	43192390	984765	Bantu
chr19	10621108	11163561	542453	Bantu
chr19	31369317	31796288	426971	Bantu
chr19	32083035	32471522	388487	Bantu
chr19	36876843	37367131	490288	Bantu
chr19	37412887	38475085	1062198	Bantu

chr19	22635052	23430435	795383	Middle East
chr19	22692592	23430435	737843	South Asia
chr19	22709636	23461423	751787	Europe
chr20	47177768	47591081	413313	Bantu
chr20	37080007	37933672	853665	Bantu
chr20	15017240	15178252	161012	Bantu
chr20	52884386	53287736	403350	Middle East
chr20	22282991	22703676	420685	Middle East
chr20	15571727	15817105	245378	Middle East
chr20	20993838	21906779	912941	Middle East
chr20	49063166	49649709	586543	Middle East
chr20	33737661	34875144	1137483	East Asia
chr20	30437522	31712589	1275067	East Asia
chr20	52804492	53288594	484102	South Asia
chr20	22194770	22708358	513588	South Asia
chr20	49097525	49653969	556444	South Asia
chr20	53581724	53816637	234913	South Asia
chr20	6893128	7164724	271596	Europe
chr20	22185332	22506049	320717	Europe
chr20	33844938	34442671	597733	Europe
chr20	8566792	8889612	322820	America
chr20	16513316	16683777	170461	America
chr21	30031030	30864811	833781	Bantu
chr21	44172913	44422295	249382	Bantu
chr21	46032094	46158737	126643	Bantu
chr21	46982335	47322335	340000	Middle East
chr21	17817942	18114357	296415	East Asia
chr21	29966286	31178625	1212339	South Asia
chr21	40326522	41030756	704234	America
chr21	41049045	41278301	229256	America
chr21	15861455	15984177	122722	America
chr22	31530730	32289918	759188	Bantu
chr22	46447097	46863377	416280	Middle East
chr22	18526789	18918824	392035	Middle East
chr22	35518987	35797832	278845	Middle East
chr22	32275353	32699672	424319	East Asia
chr22	36478027	36935795	457768	East Asia
chr22	46447097	46863377	416280	South Asia
chr22	28244191	29198150	953959	South Asia
chr22	25619339	25950749	331410	South Asia
chr22	35519850	35795412	275562	Europe
chr22	49912406	50491712	579306	Europe
chr22	30654464	30921370	266906	Europe
chr22	49855674	50278567	422893	America
chrX	121993027	122366636	373609	Bantu
chrX	126640531	127637654	997123	Bantu
chrX	109096770	110400647	1303877	Bantu

chrX	34986964	35753528	766564	Bantu
chrX	65922033	67286067	1364034	Bantu
chrX	137952577	138233542	280965	Bantu
chrX	34225721	36560061	2334340	Middle East
chrX	109689152	111504315	1815163	Middle East
chrX	98117578	99486285	1368707	Middle East
chrX	97004924	97760765	755841	Middle East
chrX	30026227	30597320	571093	Middle East
chrX	18859525	20417949	1558424	Middle East
chrX	5728176	6134767	406591	Middle East
chrX	153189819	153627145	437326	Middle East
chrX	67177760	68076640	898880	East Asia
chrX	97108437	97837882	729445	East Asia
chrX	14035798	15163507	1127709	East Asia
chrX	150438543	150901389	462846	East Asia
chrX	64405169	68053507	3648338	South Asia
chrX	20522619	21802998	1280379	South Asia
chrX	108966425	111694354	2727929	South Asia
chrX	30272489	30546430	273941	South Asia
chrX	121222635	121942321	719686	South Asia
chrX	125843757	127823996	1980239	Europe
chrX	95072450	95872001	799551	Europe
chrX	98224405	99486285	1261880	Europe
chrX	65435257	67923735	2488478	Europe
chrX	106736820	111555208	4818388	America
chrX	5650083	6135610	485527	America

Supplementary table 2.4: Lists, for each cell type, the mean methylation of CGIs inside HIRs (with its standard error), the mean methylation of CGIs localized outside these regions (with its standard error), the number of CGIs inside HIRs, the number of CGIs localized outside HIRs and the Bootstrap p-values.

Cell ID	Cell type	HIR CGIs mean	HIR SE	Other CGIs mean	Other SE	No. of HIR CGIs	No. of other CGIs	Bootstrap p-value
Hek293	cancer	18.07209546	0.74963397	21.71357874	0.27182048	1662	15159	< 1.0E-04
MCF-7	cancer	28.3134172	0.90845651	33.25740612	0.31383639	1784	16415	< 1.0E-04
Hepg2	cancer	23.58733521	0.81826007	27.39282804	0.28077965	1764	16427	1.0E-04
Cmk	cancer	26.3994764	0.89198779	33.63190574	0.31407055	1741	16102	< 1.0E-04
NB4	cancer	27.10041988	0.86443758	31.93859768	0.29668278	1745	16100	< 1.0E-04
NT2-D1	cancer	12.67613946	0.70799264	16.301809	0.26401909	1616	14855	< 1.0E-04
Gm19239	EBV	13.61689916	0.64389488	17.88822672	0.24354854	1635	15223	< 1.0E-04
Gm19240	EBV	17.0858443	0.67483608	20.74296261	0.24533451	1811	16766	< 1.0E-04
Ag04449	normal	8.291692217	0.42508266	10.14338957	0.16246635	1696	15412	1.0E-04
Ag04450	normal	11.03531603	0.56654663	13.94272058	0.21097382	1758	16285	< 1.0E-04
Ag09309	normal	13.68236452	0.59444949	17.1290823	0.22176914	1758	16355	< 1.0E-04
Ag09319	normal	11.61979758	0.59526593	14.74437777	0.22517808	1691	15459	< 1.0E-04
Ag10803	normal	12.37297695	0.58687882	16.00796514	0.22029061	1855	17306	< 1.0E-04
Fibrobl	normal	14.11595976	0.630049	17.49915036	0.23169448	1701	15907	< 1.0E-04
HAEpiC	normal	10.42137587	0.56003811	14.11197435	0.21809566	1764	16270	< 1.0E-04
HCF	normal	9.07611084	0.55552218	12.9804742	0.22368603	1591	14546	< 1.0E-04
HCM	normal	10.40447994	0.56516963	13.60023807	0.21422817	1817	16829	< 1.0E-04
HEEpiC	normal	10.04693673	0.54926804	13.38862614	0.21297418	1729	15915	< 1.0E-04
HIPEpiC	normal	10.41191083	0.5493907	13.52147836	0.21222731	1742	15980	< 1.0E-04
HMEC	normal	12.72713219	0.61298982	16.35859901	0.23280776	1736	16089	< 1.0E-04
HNPCEpiC	normal	10.14362759	0.54589589	13.08930351	0.20736638	1801	16630	< 1.0E-04
HRCEpiC	normal	9.108447913	0.56073348	12.00655231	0.21357911	1646	15161	< 1.0E-04
HSMMtube	normal	18.00726159	0.66202673	21.25326949	0.2386281	1789	16680	< 1.0E-04
NHBE	normal	11.02629679	0.57314939	14.52355975	0.22030469	1767	16389	< 1.0E-04
Skmc	normal	10.79080888	0.59161788	14.86922907	0.22582651	1741	16221	< 1.0E-04

Supplementary table 2.5: Lists, for each 5SLR identified, the chromosome, the start position, the end position and the total length. Genomic coordinates refer to assembly GRCh37/hg19.

Regions	Start	End	Length
chr1	29250635	29582124	331489
chr1	46026531	46214183	187652
chr1	63862069	64092146	230077
chr1	114057822	114381300	323478
chr1	115116433	115378793	262360
chr1	46554517	46718374	163857
chr1	51008171	51235816	227645
chr1	78052717	78289590	236873
chr1	94639336	94839130	199794
chr1	97229888	97509581	279693
chr1	98053562	98445381	391819
chr1	151286575	151611883	325308
chr1	199900613	200030369	129756
chr1	210136341	210369242	232901
chr1	243662053	243895373	233320
chr1	85066633	85177704	111071
chr2	63927885	64188651	260766
chr2	74448828	74715238	266410
chr2	99573995	99832771	258776
chr2	103756037	103972017	215980
chr2	144697382	145176389	479007
chr2	145592590	145791678	199088
chr2	148547241	149002548	455307
chr2	176927165	177270516	343351
chr2	205421745	205651824	230079
chr2	15412979	15586646	173667
chr2	22662161	22828842	166681
chr2	31733656	31971218	237562
chr2	32718848	32917771	198923
chr2	37095793	37258412	162619
chr2	43411503	43747885	336382
chr2	57948185	58116477	168292
chr2	61424903	61742537	317634
chr2	62981797	63289808	308011
chr2	72690520	72852907	162387
chr2	73549924	73794209	244285
chr2	99185422	99478291	292869
chr2	123575222	123715501	140279
chr2	135845130	136125165	280035
chr2	142960140	143128724	168584
chr2	146134867	146337410	202543
chr2	152247420	152435284	187864
chr2	155700922	155938213	237291
chr2	187897747	188105218	207471
chr2	232835431	233180541	345110
chr3	49097085	49399863	302778
chr3	94365581	94623235	257654
chr3	95238967	95423773	184806
chr3	99494533	99834482	339949
chr3	136208787	136429395	220608

chr3	13969054	14217300	248246
chr3	25797576	25976557	178981
chr3	44325263	44663213	337950
chr3	48613019	48895220	282201
chr3	58841784	59137547	295763
chr3	71400071	71615018	214947
chr3	79204857	79537773	332916
chr3	110619507	110904347	284840
chr3	114684268	114906146	221878
chr3	119975093	120114914	139821
chr3	130635092	130860757	225665
chr3	155761633	156114898	353265
chr3	156493013	156744033	251020
chr3	159979892	160278818	298926
chr3	176706064	176840396	134332
chr3	180964991	181123818	158827
chr3	181695505	181910529	215024
chr4	128619572	128993304	373732
chr4	34010143	34247945	237802
chr4	46173112	46557143	384031
chr4	67417795	67622971	205176
chr4	74983811	75148509	164698
chr4	81177928	81697216	519288
chr4	98214952	98529524	314572
chr4	123960051	124100556	140505
chr4	172431101	172637662	206561
chr5	109035495	109234306	198811
chr5	19608621	19981207	372586
chr5	43889881	44105084	215203
chr5	45050586	45393754	343168
chr5	61682074	61960620	278546
chr5	93050452	93551028	500576
chr5	114116640	114350704	234064
chr5	131013414	131263355	249941
chr5	17921280	18056522	135242
chr5	27016109	27151405	135296
chr5	37380577	37649448	268871
chr5	71456746	71647477	190731
chr5	81284883	81637960	353077
chr5	86549391	86844673	295282
chr5	126999041	127242725	243684
chr5	160787118	161026677	239559
chr6	79526516	79820785	294269
chr6	126668173	126989092	320919
chr6	132102752	132196457	93705
chr6	140338210	140976317	638107
chr6	45332304	45597525	265221
chr6	49788953	50101371	312418
chr6	84298676	84482187	183511
chr6	90706068	90847204	141136
chr6	98985278	99143313	158035
chr6	128110771	128305888	195117
chr7	69025009	69636926	611917
chr7	98521518	98713281	191763
chr7	132988393	133189530	201137

chr7	18379007	18586307	207300
chr7	40050109	40192009	141900
chr7	41571216	41871572	300356
chr7	43605654	43888375	282721
chr7	48654708	48827838	173130
chr7	93759190	93985638	226448
chr7	107010512	107177005	166493
chr7	114411347	114629622	218275
chr7	121976180	122495427	519247
chr7	127178538	127662335	483797
chr8	34598254	34831202	232948
chr8	49252900	49477598	224698
chr8	53439456	53638576	199120
chr8	64868710	65123925	255215
chr8	116359583	116569557	209974
chr8	19385418	19497670	112252
chr8	28773614	29154132	380518
chr8	35772643	36067503	294860
chr8	47708415	48104073	395658
chr8	49594702	49828725	234023
chr8	58427937	58603646	175709
chr8	63732189	63873882	141693
chr8	66307235	66408755	101520
chr8	71004379	71195443	191064
chr8	92653592	92906609	253017
chr8	99633644	100014611	380967
chr8	100117902	100308266	190364
chr8	127090898	127173121	82223
chr9	125718241	126026082	307841
chr9	84794785	85035221	240436
chr9	37785698	37927533	141835
chr9	102334727	102629879	295152
chr10	62673306	62985661	312355
chr10	74413470	74667147	253677
chr10	103706040	104007262	301222
chr10	9610081	9746948	136867
chr10	32879965	33141717	261752
chr10	50674530	50815585	141055
chr10	60345768	60592816	247048
chr10	75275890	75476096	200206
chr10	83346626	83724563	377937
chr10	106713807	106848420	134613
chr10	24907707	25031090	123383
chr11	31061804	31670889	609085
chr11	27485795	27721022	235227
chr11	30644423	31036216	391793
chr11	41655528	41784303	128775
chr11	45325027	45512886	187859
chr11	45746288	45911305	165017
chr11	55023837	55441682	417845
chr11	55647914	55896015	248101
chr11	56627239	56896648	269409
chr11	57008535	57100848	92313
chr11	59857388	60041190	183802
chr11	71895114	72237309	342195

chr11	72517367	72762960	245593
chr11	95893439	96227949	334510
chr11	108543742	108800455	256713
chr12	89009747	89366763	357016
chr12	15405896	15572168	166272
chr12	16884746	17099885	215139
chr12	65795003	65925191	130188
chr12	87278562	87505777	227215
chr12	90247389	90498984	251595
chr13	20534301	20657647	123346
chr13	51944276	52082437	138161
chr13	60540445	60789907	249462
chr13	68623480	68924857	301377
chr13	84132000	84421285	289285
chr14	75426604	75646308	219704
chr14	59676151	59822731	146580
chr14	49861650	50025588	163938
chr14	68276249	68463491	187242
chr14	71750112	72211600	461488
chr14	83608854	83776600	167746
chr14	101178423	101347971	169548
chr14	105636111	105757770	121659
chr15	72142918	72396120	253202
chr15	49675957	49862449	186492
chr15	64593136	64952886	359750
chr15	84016770	84224384	207614
chr15	84319667	84496111	176444
chr16	34269791	34737702	467911
chr16	61845083	62098061	252978
chr16	63811605	64117746	306141
chr16	34917709	35149023	231314
chr16	46805547	47144825	339278
chr17	28223981	28512879	288898
chr17	27410713	27594951	184238
chr17	30947344	31213911	266567
chr17	33121487	33398915	277428
chr17	58271331	58508582	237251
chr17	62469023	62706874	237851
chr17	67929661	68017584	87923
chr18	34395780	34824354	428574
chr18	31557831	31867235	309404
chr18	41300552	41565496	264944
chr19	10931683	11147333	215650
chr20	33474216	33644358	170142
chr20	11423194	11555408	132214
chr20	13388946	13660066	271120
chr20	30247858	30480632	232774
chr20	34425967	34610890	184923
chr20	32988114	33235384	247270
chr21	38658252	38867218	208966
chr22	28575508	28914746	339238
chrX	63515628	63803599	287971

Supplementary table 2.6: Lists, for each cell type, the mean methylation of CGIs inside 5SLRs (with its standard error), the mean methylation of CGIs localized outside these regions (with its standard error), the number of CGIs inside 5SLRs, the number of CGIs localized outside 5SLRs and the Bootstrap p-values.

Cell ID	Cell type	5SLR CGIs mean	5SLR SE	Other CGIs mean	Other SE	No. of 5SLR CGIs	No. of other CGIs	Bootstrap p-value
Hek293	cancer	14.13274539	1.8511369	21.45874374	0.25828709	241	16580	1.0E-04
MCF-7	cancer	27.71896703	2.41106012	32.84600737	0.29918481	260	17939	0.0177
Hepg2	cancer	22.04356073	2.04318492	27.0943291	0.26797811	254	17937	0.0115
Cmk	cancer	25.47229476	2.31296117	33.0334249	0.29927342	253	17590	0.0014
NB4	cancer	25.0108622	2.26837969	31.55868889	0.28301359	254	17591	0.0022
NT2-D1	cancer	11.17169623	1.67069663	16.01758022	0.25061011	243	16228	0.0069
Gm19239	EBV	10.98331376	1.41763614	17.56968601	0.2311772	245	16613	1.0E-04
Gm19240	EBV	11.81273641	1.36922079	20.50909299	0.23348113	262	18315	<1.0E-04
Ag04449	normal	7.649343629	0.98891071	9.995338002	0.15394158	259	16849	0.0256
Ag04450	normal	7.43814311	1.07195526	13.75182071	0.20059277	264	17779	<1.0E-04
Ag09309	normal	12.56376747	1.470223	16.85736998	0.21045972	265	17848	0.0043
Ag09319	normal	8.559445713	1.25525234	14.52322794	0.21365955	250	16900	1.0E-04
Ag10803	normal	9.817678077	1.30871342	15.74044319	0.20913789	273	18888	2.0E-04
Fibrobl	normal	11.861235	1.46617254	17.24912499	0.22019253	251	17357	0.0013
HAEpiC	normal	7.342581647	1.13871897	13.84691597	0.20668094	266	17768	1.0E-04
HCF	normal	7.011662842	1.220579	12.67769232	0.21138466	234	15903	3.0E-04
HCM	normal	6.637513291	1.11430818	13.38838552	0.20339969	275	18371	<1.0E-04
HEEpiC	normal	8.023852996	1.22452669	13.13591306	0.20173261	258	17386	3.0E-04
HIPEpiC	normal	7.311095836	1.14318102	13.3044253	0.20114004	262	17460	<1.0E-04
HMEC	normal	9.655115391	1.28904814	16.09928353	0.22093885	261	17564	<1.0E-04
HNPCEpiC	normal	6.66867062	1.03489542	12.89298353	0.19688426	271	18160	<1.0E-04
HRCEpiC	normal	6.161202927	1.0698717	11.80499639	0.20272315	245	16562	<1.0E-04
HSMMtube	normal	13.95838672	1.38350565	21.03812762	0.22722535	259	18210	<1.0E-04
NHBE	normal	8.672499068	1.25910838	14.26513149	0.20885555	266	17890	2.0E-04
Skmc	normal	7.906367112	1.17897318	14.57038231	0.21436746	260	17702	<1.0E-04

Supplementary table 2.7: Lists, for each cell type, the mean methylation of CGIs containing CEs (with its standard error), the mean methylation of CGIs that not contain CEs (with its standard error), the number of CE CGIs, the number of non-CE CGIs and the Bootstrap pvalues.

Cell ID	Cell type	CE CGIs mean	CE SE	Other CGIs mean	Other SE	No. of CE CGIs	No. of other CGIs	Bootstrap p-value
Hek293	cancer	18.06194902	0.30389391	26.63876414	0.4472349	10365	6456	<1.0E-04
MCF-7	cancer	29.31174071	0.3666015	38.08065182	0.4934211	11016	7183	<1.0E-04
Hepg2	cancer	24.10220364	0.32920185	31.47347397	0.44026143	10981	7210	<1.0E-04
Cmk	cancer	30.52155783	0.37046424	36.63095243	0.49000397	10820	7023	<1.0E-04
NB4	cancer	28.67202877	0.3463578	35.77711741	0.46939085	10829	7016	<1.0E-04
NT2-D1	cancer	12.66381809	0.28249896	21.215268	0.45310448	10149	6322	<1.0E-04
Gm19239	EBV	14.64260305	0.2693243	21.94428629	0.40303282	10321	6537	<1.0E-04
Gm19240	EBV	17.38779645	0.27665389	24.94318509	0.39712404	11204	7373	<1.0E-04
Ag04449	normal	8.162944221	0.17489162	12.76902457	0.27540567	10434	6674	<1.0E-04
Ag04450	normal	11.42000275	0.23610187	17.08627404	0.34446075	10912	7131	<1.0E-04
Ag09309	normal	14.45463991	0.25464626	20.3435999	0.35121572	10916	7197	<1.0E-04
Ag09319	normal	11.9928598	0.24871994	18.27074267	0.37295181	10475	6675	<1.0E-04
Ag10803	normal	13.05760133	0.25026711	19.48613818	0.35094952	11416	7745	<1.0E-04
Fibrobl	normal	14.65374098	0.26273602	21.06414299	0.37384171	10690	6918	<1.0E-04
HAEpiC	normal	11.11014802	0.23744364	17.78682268	0.36280708	10901	7133	<1.0E-04
HCF	normal	10.25670137	0.24188447	16.36772469	0.3777372	9961	6176	<1.0E-04
HCM	normal	11.02712334	0.2386924	16.71753014	0.35013204	11235	7411	<1.0E-04
HEEpiC	normal	10.74731068	0.23342271	16.64781495	0.35385824	10725	6919	<1.0E-04
HIPEpiC	normal	10.83446026	0.2315216	16.90413784	0.35407515	10769	6953	<1.0E-04
HMEC	normal	13.37336375	0.25791805	20.05535694	0.38294662	10805	7020	<1.0E-04
HNPCEpiC	normal	10.51655713	0.22922112	16.27207527	0.34130963	11114	7317	<1.0E-04
HRCEpiC	normal	9.776230222	0.23552555	14.82870027	0.35643076	10332	6475	<1.0E-04
HSMMtube	normal	17.72691312	0.26767559	25.78575942	0.38746289	11108	7361	<1.0E-04
NHBE	normal	11.61694328	0.24213548	18.12160123	0.36439618	10993	7163	<1.0E-04
Skmc	normal	11.82268113	0.25106864	18.53088857	0.36932762	10863	7099	<1.0E-04

Supplementary table 2.8: Lists, for each cell type, the mean methylation of HIR+CE CGIs (with its standard error), the mean methylation of CE CGIs (with its standard error), the number of HIR+CE CGIs, the number of CE CGIs and the Bootstrap p-values.

Cell ID	Cell type	HIR+CE CGIs mean	HIR+CE SE	CE CGIs mean	CE SE	No. of HIR+CE CGIs	No. of CE CGIs	Bootstrap p-value
Hek293	cancer	15.83623485	0.89925163	18.30541256	0.32238172	1022	9343	0.0069
MCF-7	cancer	24.78486437	1.10807098	29.80379249	0.38789151	1080	9936	< 1.E-04
Hepg2	cancer	20.94443157	1.00181754	24.44453201	0.34819505	1074	9907	0.0011
Cmk	cancer	24.43245785	1.10632827	31.18149102	0.39215476	1058	9762	< 1.E-04
NB4	cancer	24.48509553	1.05010054	29.12728863	0.36637262	1062	9767	< 1.E-04
NT2-D1	cancer	9.818266859	0.78875638	12.97139486	0.30103731	990	9159	7.E-04
Gm19239	EBV	11.78969143	0.764194	14.9486766	0.28655855	1000	9321	1.E-04
Gm19240	EBV	14.98866755	0.80971263	17.64898427	0.29372972	1100	10104	1.E-03
Ag04449	normal	7.017109602	0.48293697	8.28858024	0.18666602	1031	9403	0.013
Ag04450	normal	9.514735893	0.68227508	11.62520914	0.25091804	1061	9851	0.0025
Ag09309	normal	12.12900686	0.73520006	14.70528151	0.27062392	1062	9854	9.E-04
Ag09319	normal	9.680989007	0.69333385	12.24551792	0.26517444	1032	9443	7.E-04
Ag10803	normal	10.67151203	0.70994591	13.31510631	0.26636641	1112	10304	7.E-04
Fibrobl	normal	12.49566357	0.76714702	14.88582598	0.27896015	1038	9652	0.0039
HAEpiC	normal	8.676708592	0.65495135	11.37417892	0.25329867	1067	9834	2.E-04
HCF	normal	7.853319139	0.6583744	10.51984736	0.25836529	983	8978	< 1.E-04
HCM	normal	8.903958885	0.6712296	11.25663199	0.25425245	1096	10139	0.0017
HEEpiC	normal	8.343485959	0.64060805	11.00764015	0.24909169	1048	9677	2.E-04
HIPEpiC	normal	8.803586043	0.63911698	11.05664923	0.24705585	1062	9707	0.0015
HMEC	normal	10.33919878	0.70639778	13.69995249	0.27518135	1050	9755	< 1.E-04
HNPCEpiC	normal	8.630212897	0.64394071	10.72105062	0.24421423	1087	10027	0.0028
HRCEpiC	normal	7.883028293	0.66014666	9.982026289	0.25099236	1013	9319	0.0027
HSMMtube	normal	15.67666393	0.7918037	17.94726892	0.28389121	1078	10030	0.0042
NHBE	normal	9.133926979	0.67076771	11.88441048	0.25816884	1069	9924	2.E-04
Skmc	normal	9.147031134	0.69973955	12.11139481	0.26756361	1058	9805	< 1.E-04

Supplementary table 2.9: Lists, for each cell type, the mean methylation of 5SLR+CE CGIs (with its standard error), the mean methylation of CE CGIs (with its standard error), the number of 5SLR+CE CGIs, the number of CE CGIs and the Bootstrap p-values.

Cell ID	Cell type	5LSR+CE CGIs mean	5LSR +CE SE	CE CGIs mean	CE SE	No. of 5LSR +CE CGIs	No. of CE CGIs	Bootstrap p-value
Hek293	cancer	10.37628825	1.90918117	18.17939079	0.30701808	156	10209	1.0E-04
MCF-7	cancer	22.93389404	2.80151076	29.40812542	0.36966261	164	10852	0.0154
Hepg2	cancer	18.90791051	2.32056151	24.1799812	0.3322738	162	10819	0.0242
Cmk	cancer	23.01581493	2.73381193	30.63492912	0.37369178	161	10659	0.0057
NB4	cancer	22.02005809	2.66749511	28.77368555	0.34920252	163	10666	0.0066
NT2-D1	cancer	5.441370892	1.1917729	12.77730129	0.28618458	157	9992	1.0E-04
Gm19239	EBV	7.712200781	1.33052292	14.74827021	0.27254793	155	10166	2.0E-04
Gm19240	EBV	8.662255749	1.29448932	17.51982193	0.27997111	167	11037	<1.0E-04
Ag04449	normal	5.637730061	0.92448072	8.20301918	0.17703586	163	10271	0.0263
Ag04450	normal	4.613737254	0.83680997	11.52707329	0.23931281	169	10743	<1.0E-04
Ag09309	normal	8.804471169	1.51644723	14.54188903	0.25743204	166	10750	0.0013
Ag09319	normal	4.964918459	0.99712684	12.09979784	0.25190454	157	10318	<1.0E-04
Ag10803	normal	7.41177632	1.36409965	13.14396579	0.25315534	172	11244	0.0017
Fibrobl	normal	9.453901503	1.66012629	14.73224962	0.26545585	159	10531	0.0053
HAEpiC	normal	5.190239125	1.12907275	11.20281034	0.24040786	168	10733	3.0E-04
HCF	normal	4.781792145	1.15844024	10.34154041	0.24488216	152	9809	3.0E-04
HCM	normal	4.342251171	1.08079501	11.13289663	0.24173463	175	11060	<1.0E-04
HEEpiC	normal	5.087948543	1.14169091	10.83410578	0.23625912	162	10563	1.0E-04
HIPEpiC	normal	4.471057319	1.01658838	10.93408535	0.23448096	166	10603	<1.0E-04
HMEC	normal	5.524053762	0.86312715	13.49433798	0.26138397	164	10641	<1.0E-04
HNPCEpiC	normal	3.873851058	0.81003083	10.62159215	0.23235585	173	10941	<1.0E-04
HRCEpiC	normal	3.569559848	0.84418476	9.872618458	0.23870101	158	10174	1.0E-04
HSMMtube	normal	10.82291589	1.39448125	17.8297319	0.27075048	163	10945	2.0E-04
NHBE	normal	5.488231834	1.1643759	11.71263362	0.24513501	169	10824	2.0E-04
Skmc	normal	5.756426782	1.16640023	11.91624366	0.25420349	165	10698	2.0E-04

Supplementary table 2.10: Lists, for each cell type, the number, the mean methylation and the standard error of 59 CGIs, intragenic CGIs, 39 CGIs and intergenic CGIs.

		5' CGIs			Intragenic CGIs			3' CGIs			Intergenic CGIs		
Cell ID	**Cell type**	**number**	**mean**	**SE**	**number**	**mean**	**SE**	**number**	**mean**	**SE**	**number**	**mean**	**SE**
Hek293	cancer	10947	12.55	0.24	995	64.95	1.2	746	55.61	1.46	3187	31.54	0.64
MCF-7	cancer	11431	21.50	0.32	1238	74.17	1	888	65.73	1.31	3580	47.31	0.69
Hepg2	cancer	11310	17.76	0.28	1291	64.87	1	966	60.73	1.26	3592	34.67	0.60
Cmk	cancer	11386	23.22	0.33	1165	70.83	1.1	843	61.89	1.38	3430	46.84	0.69
NB4	cancer	11343	21.75	0.31	1145	70.30	1	856	60.96	1.33	3473	44.21	0.65
NT2-D1	cancer	10744	8.73	0.22	926	57.94	1.4	723	44.37	1.63	3112	22.58	0.65
Gm19239	EBV	10683	9.92	0.21	1110	57.05	1.2	821	45.17	1.38	3271	22.63	0.54
Gm19240	EBV	11578	12.22	0.22	1314	60.63	1.1	945	48.41	1.31	3683	25.71	0.54
Ag04449	normal	11105	5.92	0.13	880	35.20	1.2	735	26.95	1.16	3377	13.33	0.37
Ag04450	normal	11411	7.75	0.18	1164	47.92	1.1	865	35.21	1.27	3554	17.01	0.47
Ag09309	normal	11365	10.04	0.2	1219	53.42	1.1	874	41.84	1.28	3613	20.44	0.46
Ag09319	normal	10969	8.16	0.19	1043	49.83	1.2	800	37.98	1.35	3348	19.19	0.51
Ag10803	normal	11886	8.66	0.19	1333	52.14	1	969	40.00	1.24	3868	19.41	0.47
FibrobI	normal	11036	9.88	0.2	1218	55.32	1.1	888	43.14	1.28	3454	21.27	0.50
HAEpiC	normal	11356	7.39	0.18	1174	48.80	1.2	872	35.81	1.31	3592	17.81	0.49
HCF	normal	10325	6.61	0.18	974	47.72	1.3	774	35.08	1.36	3140	16.39	0.51
HCM	normal	11704	7.09	0.18	1218	48.23	1.2	918	35.37	1.27	3736	16.73	0.48
HEEpiC	normal	11181	7.12	0.18	1143	47.11	1.2	840	33.88	1.28	3462	16.74	0.49
HIPEpiC	normal	11280	7.46	0.18	1105	47.52	1.2	840	34.57	1.3	3480	16.71	0.48
HMEC	normal	11259	9.10	0.2	1156	52.98	1.2	867	40.72	1.34	3514	20.88	0.52
HNPCEpiC	normal	11625	6.95	0.17	1174	46.57	1.2	878	33.32	1.28	3676	16.12	0.46
HRCEpiC	normal	10767	6.27	0.18	1026	45.96	1.3	780	33.33	1.36	3260	14.78	0.48
HSMMtube	normal	11506	12.90	0.21	1307	58.57	1	938	47.63	1.23	3658	26.84	0.53
NHBE	normal	11450	7.68	0.19	1182	49.57	1.2	890	37.16	1.28	3587	18.47	0.50
Skmc	normal	11216	7.44	0.19	1238	51.56	1.1	915	39.05	1.26	3572	18.44	0.50

For each cell line, Kruskal Wallis Test, p-value $\leq 2.2\ 10^{-16}$

Supplementary table 2.11: Lists, for each cell type and for each CGIs class (59 CGIs, intragenic CGIs, 39 CGIs and intergenic CGIs) the number and the mean methylation of CGIs containing CEs (with its standard error), the number and the mean methylation of CGIs that do not contain CEs (with its standard error), and the Bootstrap p-values.

		5 ' CGIs							Intragenic CGIs							3' CGIs							Intergenic CGIs						
		CE CGIs			Other CGIs				CE CGIs			Other CGIs				CE CGIs			Other CGIs				CE CGIs			Other CGIs			
Cell ID	Cell type	n.	mean	SE	n.	mean	SE	p-value	n.	mean	SE	n.	mean	SE	p-value	n.	mean	SE	n.	mean	SE	p-value	n.	mean	SE	n.	mean	SE	p-value
Hek293	cancer	7117	11.31	0.28	3831	14.85	0.45	$< 10^{-4}$	526	64.80	1.65	469	65.11	1.72	0.4456	445	57.70	1.87	301	52.53	2.34	0.9589	1228	23.56	0.94	1960	36.52	0.84	$< 10^{-4}$
MCF-7	cancer	7376	20.58	0.38	4056	23.18	0.56	$< 10^{-4}$	646	74.77	1.36	592	73.52	1.44	0.7359	518	67.43	1.70	370	63.34	2.05	0.9417	1306	40.35	1.11	2275	51.30	0.87	$< 10^{-4}$
Hepg2	cancer	7302	16.61	0.34	4009	19.84	0.52	$< 10^{-4}$	680	66.66	1.40	611	62.89	1.50	0.9700	559	62.73	1.64	407	57.98	1.97	0.9692	1300	27.76	0.94	2293	38.57	0.75	$< 10^{-4}$
Cmk	cancer	7338	22.94	0.40	4049	23.71	0.57	0.1354	617	73.28	1.41	548	68.08	1.58	0.9924	488	64.18	1.77	355	58.73	2.17	0.974	1261	41.65	1.11	2170	49.84	0.88	$< 10^{-4}$
NB4	cancer	7325	21.23	0.37	4019	22.70	0.53	0.0100	605	70.85	1.40	540	69.68	1.51	0.7146	496	63.82	1.67	360	57.01	2.15	0.9952	1274	37.67	1.01	2200	47.99	0.83	$< 10^{-4}$
NT2-D1	cancer	6995	7.71	0.26	3750	10.61	0.42	$< 10^{-4}$	484	55.41	1.99	442	60.71	2.02	0.0332	425	45.31	2.12	298	43.04	2.54	0.7574	1198	13.15	0.83	1915	28.47	0.90	$< 10^{-4}$
Gm19239	EBV	6964	8.99	0.24	3720	11.65	0.40	$< 10^{-4}$	587	56.50	1.61	523	57.67	1.66	0.3053	484	45.70	1.80	337	44.41	2.17	0.6844	1225	14.73	0.72	2047	27.35	0.73	$< 10^{-4}$
Gm19240	EBV	7469	11.11	0.26	4110	14.23	0.41	$< 10^{-4}$	698	60.39	1.47	616	60.90	1.51	0.4053	547	49.79	1.73	398	46.52	2.00	0.8881	1334	17.31	0.74	2350	30.46	0.71	$< 10^{-4}$
Ag04449	normal	7193	5.37	0.15	3913	6.93	0.25	$< 10^{-4}$	448	32.50	1.63	432	38.00	1.64	0.0103	425	28.03	1.56	310	25.48	1.74	0.8632	1274	8.83	0.49	2104	16.04	0.51	$< 10^{-4}$
Ag04450	normal	7359	6.97	0.21	4053	9.16	0.34	$< 10^{-4}$	617	46.96	1.61	547	49.01	1.63	0.1848	499	37.09	1.71	366	32.66	1.91	0.9513	1304	11.11	0.63	2251	20.43	0.63	$< 10^{-4}$
Ag09309	normal	7339	9.29	0.24	4027	11.40	0.35	$< 10^{-4}$	623	52.57	1.56	596	54.31	1.51	0.2084	511	43.15	1.69	363	39.98	1.96	0.8907	1303	14.51	0.68	2311	23.78	0.60	$< 10^{-4}$
Ag09319	normal	7117	7.27	0.22	3853	9.81	0.36	$< 10^{-4}$	553	48.76	1.74	490	51.05	1.79	0.1807	463	39.38	1.80	337	36.06	2.05	0.8828	1258	12.46	0.68	2091	23.24	0.69	$< 10^{-4}$
Ag10803	normal	7604	7.77	0.23	4283	10.24	0.35	$< 10^{-4}$	693	51.19	1.49	640	53.17	1.46	0.1695	553	41.69	1.65	416	37.75	1.86	0.9441	1367	12.30	0.66	2502	23.30	0.62	$< 10^{-4}$
FibrobI	normal	7156	9.05	0.24	3881	11.39	0.37	$< 10^{-4}$	639	54.80	1.53	579	55.89	1.53	0.3147	517	44.02	1.69	371	41.90	1.96	0.797	1267	14.43	0.70	2188	25.22	0.66	$< 10^{-4}$
HAEpiC	normal	7331	6.62	0.21	4026	8.78	0.34	$< 10^{-4}$	615	47.10	1.65	559	50.66	1.68	0.0676	510	36.50	1.72	362	34.85	2.02	0.7205	1298	10.05	0.62	2295	22.19	0.67	$< 10^{-4}$
HCF	normal	6773	5.98	0.21	3553	7.81	0.34	$< 10^{-4}$	522	47.48	1.80	452	48.00	1.85	0.4240	450	34.95	1.77	324	35.26	2.13	0.4675	1196	9.52	0.64	1945	20.60	0.71	$< 10^{-4}$
HCM	normal	7522	6.37	0.21	4183	8.37	0.33	$< 10^{-4}$	648	47.29	1.61	570	49.31	1.66	0.1872	529	36.96	1.69	389	33.20	1.92	0.9281	1348	9.84	0.63	2389	20.62	0.65	$< 10^{-4}$
HEEpiC	normal	7232	6.43	0.21	3950	8.39	0.33	$< 10^{-4}$	608	45.70	1.63	535	48.71	1.70	0.0922	488	34.05	1.69	352	33.65	1.96	0.562	1288	10.49	0.64	2175	20.43	0.66	$< 10^{-4}$
HIPEpiC	normal	7282	6.54	0.20	3999	9.13	0.34	$< 10^{-4}$	581	45.88	1.68	524	49.33	1.72	0.0755	484	35.32	1.73	356	33.55	1.95	0.7564	1296	10.26	0.61	2185	20.53	0.65	$< 10^{-4}$
HMEC	normal	7276	8.23	0.24	3984	10.70	0.38	$< 10^{-4}$	615	52.15	1.61	541	53.92	1.71	0.2315	503	41.27	1.75	364	39.97	2.07	0.6856	1296	13.84	0.71	2219	24.98	0.70	$< 10^{-4}$
HNPCEpiC	normal	7485	6.33	0.20	4141	8.07	0.32	$< 10^{-4}$	618	46.01	1.63	556	47.19	1.67	0.3054	505	34.63	1.71	373	31.55	1.91	0.8868	1337	9.32	0.59	2340	20.00	0.63	$< 10^{-4}$
HRCEpiC	normal	7009	5.59	0.21	3759	7.53	0.34	$< 10^{-4}$	549	45.54	1.74	477	46.45	1.83	0.3611	456	34.75	1.81	324	31.33	2.07	0.891	1243	8.68	0.60	2018	18.53	0.67	$< 10^{-4}$
HSMMtube	normal	7388	11.56	0.25	4119	15.29	0.40	$< 10^{-4}$	689	57.60	1.45	618	59.65	1.48	0.1625	540	49.25	1.61	398	45.43	1.88	0.9369	1319	18.12	0.73	2340	31.75	0.71	$< 10^{-4}$
NHBE	normal	7392	6.93	0.22	4059	9.04	0.34	$< 10^{-4}$	628	48.36	1.62	554	50.94	1.68	0.1309	513	37.23	1.70	377	37.07	1.96	0.5316	1312	11.29	0.66	2276	22.60	0.68	$< 10^{-4}$
Skmc	normal	7259	6.72	0.23	3958	8.74	0.35	$< 10^{-4}$	648	50.51	1.57	590	52.71	1.56	0.1601	525	39.89	1.67	390	37.92	1.93	0.784	1300	10.80	0.66	2273	22.81	0.68	$< 10^{-4}$

Supplementary table 3.1: Characteristics of cell lines used in this study

Cell line	Karyotype	Lineage	Tissue	Karyotype	Sex
HUVEC	normal	mesoderm	blood vessel	normal	U
Monocytes-CD14+-RO01746 (CD14+)	normal	mesoderm	monocytes	normal	F
HMEC	normal	ectoderm	breast	normal	U
HSMM	normal	mesoderm	muscle	normal	U
HSMMtube	normal	mesoderm	muscle	normal	U
NH-A	normal	ectoderm	brain	normal	U
NHDF-Ad	normal	mesoderm	skin	normal	F
NHEK	normal	ectoderm	skin	normal	U
NHLF	normal	endoderm	lung	normal	U
K562	cancer	mesoderm	blood	cancer	F
HeLa-S3	cancer	ectoderm	cervix	cancer	F
HepG2	cancer	endoderm	liver	cancer	M
Dnd41	cancer	mesoderm	blood	cancer	M

Supplementary table 3.2: Raw data used to calculate CGIs enriched with H3K4me3, H3K27ac and H3K36me3 in different cell lines.

Histone mark	Signature of selective pressure	Cell lines	Category	No.of CGIs with peaks under selective pressure (k)	No.of CGIs under selective pressure (n)	No. of CGIs with peaks (M)	Total no. of CGIs (N)	Hypergeo-metric P-value	bonferroni P-value
H3K4me3	**HIR**	HUVEC	normal	1652	2545	15934	27718	4.45E-16	6.7E-15
		CD14+	normal	1652	2545	16121	27718	1.19E-13	1.8E-12
		HMEC	normal	1666	2545	16108	27718	7.47E-16	1.1E-14
		HSMM	normal	1766	2545	17071	27718	3.03E-18	4.5E-17
		HSMMtube	normal	1657	2545	16263	27718	1.31E-12	2.0E-11
		NH-A	normal	1677	2545	16418	27718	1.89E-13	2.8E-12
		NHDF-Ad	normal	1660	2545	15914	27718	1.32E-17	2.0E-16
		NHEK	normal	1715	2545	16659	27718	7.05E-16	1.1E-14
		NHLF	normal	1657	2545	15852	27718	5.22E-18	7.8E-17
		K562	cancer	1422	2545	13719	27718	5.97E-12	9.0E-11
		HeLa-S3	cancer	1418	2545	13799	27718	1.41E-10	2.1E-09
		HepG2	cancer	1616	2545	15715	27718	1.08E-13	1.6E-12
		Dnd41	cancer	1247	2545	11643	27718	3.66E-14	5.5E-13
	5LSR	HUVEC	normal	256	348	15934	27718	1.27E-10	1.9E-09
		CD14+	normal	247	348	16121	27718	2.42E-07	3.6E-06
		HMEC	normal	243	348	16108	27718	2.19E-06	3.3E-05
		HSMM	normal	270	348	17071	27718	4.94E-11	7.4E-10
		HSMMtube	normal	261	348	16263	27718	4.85E-11	7.3E-10
		NH-A	normal	262	348	16418	27718	8.69E-11	1.3E-09
		NHDF-Ad	normal	260	348	15914	27718	4.60E-12	6.9E-11
		NHEK	normal	253	348	16659	27718	2.83E-07	4.3E-06
		NHLF	normal	242	348	15852	27718	7.07E-07	1.1E-05
		K562	cancer	207	348	13719	27718	6.84E-05	1.0E-03
		HeLa-S3	cancer	195	348	13799	27718	0.008125	0.1218699
		HepG2	cancer	234	348	15715	27718	2.04E-05	3.1E-04
		Dnd41	cancer	174	348	11643	27718	0.001043	0.0156501
	CE	HUVEC	normal	9032	13288	15934	27718	1.57E-254	2.4E-253
		CD14+	normal	8941	13288	16121	27718	7.02E-194	1.1E-192
		HMEC	normal	9173	13288	16108	27718	3.47E-277	5.2E-276
		HSMM	normal	9604	13288	17071	27718	8.88E-274	1.3E-272
		HSMMtube	normal	9270	13288	16263	27718	4.04E-287	6.1E-286
		NH-A	normal	9330	13288	16418	27718	8.18E-283	1.2E-281
		NHDF-Ad	normal	8975	13288	15914	27718	2.02E-237	3.0E-236
		NHEK	normal	9439	13288	16659	27718	2.21E-282	3.3E-281
		NHLF	normal	8992	13288	15852	27718	8.48E-254	1.3E-252
		K562	cancer	7700	13288	13719	27718	8.67E-162	1.3E-160
		HeLa-S3	cancer	7874	13288	13799	27718	7.91E-203	1.2E-201
		HepG2	cancer	8877	13288	15715	27718	1.81E-235	2.7E-234
		Dnd41	cancer	6669	13288	11643	27718	1.54E-155	2.3E-154
H3K27ac	**HIR**	HUVEC	normal	1290	2545	12091	27718	2.45E-14	3.7E-13
		CD14+	normal	1253	2545	12082	27718	8.35E-10	1.3E-08
		HMEC	normal	1226	2545	11595	27718	5.53E-12	8.3E-11
		HSMM	normal	1212	2545	11658	27718	1.25E-09	1.9E-08
		HSMMtube	normal	1178	2545	11452	27718	4.61E-08	6.9E-07
		NH-A	normal	1267	2545	11935	27718	3.47E-13	5.2E-12
		NHDF-Ad	normal	1328	2545	12985	27718	7.02E-09	1.1E-07
		NHEK	normal	1234	2545	11873	27718	7.50E-10	1.1E-08
		NHLF	normal	1308	2545	12452	27718	2.80E-12	4.2E-11
		K562	cancer	1271	2545	11652	27718	1.47E-17	2.2E-16

		HeLa-S3	cancer	1278	2545	11893	27718	3.03E-15	4.6E-14
		HepG2	cancer	1139	2545	10681	27718	8.13E-12	1.2E-10
		Dnd41	cancer	1299	2545	12392	27718	7.60E-12	1.1E-10
	5LSR	HUVEC	normal	202	348	12091	27718	2.05E-08	3.1E-07
		CD14+	normal	194	348	12082	27718	1.80E-06	2.7E-05
		HMEC	normal	185	348	11595	27718	7.36E-06	1.1E-04
		HSMM	normal	191	348	11658	27718	4.94E-07	7.4E-06
		HSMMtube	normal	185	348	11452	27718	2.93E-06	4.4E-05
		NH-A	normal	198	348	11935	27718	6.89E-08	1.0E-06
		NHDF-Ad	normal	221	348	12985	27718	1.18E-10	1.8E-09
		NHEK	normal	186	348	11873	27718	2.49E-05	3.7E-04
		NHLF	normal	193	348	12452	27718	2.94E-05	4.4E-04
		K562	cancer	176	348	11652	27718	0.000517	7.7E-03
		HeLa-S3	cancer	178	348	11893	27718	0.000776	0.01
		HepG2	cancer	174	348	10681	27718	4.91E-06	7.36E-05
		Dnd41	cancer	182	348	12392	27718	0.001797	0.03
	CE	HUVEC	normal	7015	13288	12091	27718	4.02E-193	6.0E-192
		CD14+	normal	6758	13288	12082	27718	5.05E-122	7.6E-121
		HMEC	normal	6793	13288	11595	27718	3.39E-200	5.1E-199
		HSMM	normal	6852	13288	11658	27718	2.85E-209	4.3E-208
		HSMMtube	normal	6751	13288	11452	27718	1.39E-209	2.1E-208
		NH-A	normal	7007	13288	11935	27718	2.34E-215	3.5E-214
		NHDF-Ad	normal	7436	13288	12985	27718	1.77E-188	2.7E-187
		NHEK	normal	6883	13288	11873	27718	2.33E-185	3.5E-184
		NHLF	normal	7177	13288	12452	27718	1.51E-188	2.3E-187
		K562	cancer	6654	13288	11652	27718	4.28E-150	6.4E-149
		HeLa-S3	cancer	6971	13288	11893	27718	2.77E-210	4.2E-209
		HepG2	cancer	6289	13288	10681	27718	2.03E-184	3.1E-183
		Dnd41	cancer	6918	13288	12392	27718	3.12E-124	4.7E-123
H3K36me3	**HIR**	HUVEC	normal	952	2545	9878	27718	0.024276	0.36
		CD14+	normal	750	2545	7791	27718	0.0524	0.79
		HMEC	normal	944	2545	9943	27718	0.085784	1
		HSMM	normal	1195	2545	12139	27718	0.000354	5.3E-03
		HSMMtube	normal	991	2545	10456	27718	0.088709	1
		NH-A	normal	858	2545	8652	27718	0.002109	0.03
		NHDF-Ad	normal	1033	2545	10426	27718	0.000558	8.4E-03
		NHEK	normal	1069	2545	11179	27718	0.034092	0.51
		NHLF	normal	1077	2545	11387	27718	0.088354	1
		K562	cancer	1059	2545	10009	27718	8.07E-10	1.2E-08
		HeLa-S3	cancer	1016	2545	9759	27718	9.75E-08	1.5E-06
		HepG2	cancer	946	2545	9488	27718	0.00051	7.6E-03
		Dnd41	cancer	1129	2545	11261	27718	2.77E-05	4.2E-04
	5LSR	HUVEC	normal	142	348	9878	27718	0.019402	0.29
		CD14+	normal	108	348	7791	27718	0.100787	1
		HMEC	normal	131	348	9943	27718	0.226059	1
		HSMM	normal	175	348	12139	27718	0.006163	0.09
		HSMMtube	normal	127	348	10456	27718	0.661453	1
		NH-A	normal	125	348	8652	27718	0.025817	0.39
		NHDF-Ad	normal	156	348	10426	27718	0.002367	0.04
		NHEK	normal	153	348	11179	27718	0.074589	1
		NHLF	normal	153	348	11387	27718	0.124186	1
		K562	cancer	158	348	10009	27718	0.000138	2.1E-03
		HeLa-S3	cancer	170	348	9759	27718	5.59E-08	8.4E-07
		HepG2	cancer	144	348	9488	27718	0.002193	0.03
		Dnd41	cancer	156	348	11261	27718	0.048921	0.73

	CE	HUVEC	normal	5709	13288	9878	27718	1.16E-132	1.7E-131
		CD14+	normal	4309	13288	7791	27718	1.34E-53	2.0E-52
		HMEC	normal	5725	13288	9943	27718	2.97E-128	4.5E-127
		HSMM	normal	6907	13288	12139	27718	4.92E-154	7.4E-153
		HSMMtube	normal	5894	13288	10456	27718	1.43E-106	2.1E-105
		NH-A	normal	5064	13288	8652	27718	1.13E-125	1.7E-124
		NHDF-Ad	normal	6111	13288	10426	27718	6.13E-169	9.2E-168
		NHEK	normal	6432	13288	11179	27718	2.83E-153	4.2E-152
		NHLF	normal	6540	13288	11387	27718	9.63E-155	1.4E-153
		K562	cancer	5778	13288	10009	27718	1.44E-133	2.2E-132
		HeLa-S3	cancer	5714	13288	9759	27718	1.09E-150	1.6E-149
		HepG2	cancer	5378	13288	9488	27718	1.22E-98	1.8E-97
		Dnd41	cancer	6373	13288	11261	27718	1.65E-126	2.5E-125

Supplementary table 3.3: Raw data used to calculate CGIs enriched with H3K4me3, H3K27ac and H3K36me3 at 5', intragenic, 3' and intergenic locations in different cell lines.

Histone mark	Signature of selective pressure	CGIs in different gene region	Cell lines	Category	No.of CGIs with peaks under selective pressure (k)	No.of CGIs under selective pressure (n)	No. of CGIs with peaks (M)	Total no. of CGIs (N)	Hypergeometric P-value	bonferroni P-value
H3K34me3	**HIR**	**5'CGIs**	HUVEC	normal	1233	1535	11399	15680	1.35E-13	1.76E-12
			CD14+	normal	1222	1535	11296	15680	3.31E-13	4.31E-12
			HMEC	normal	1237	1535	11565	15680	1.82E-11	2.36E-10
			HSMM	normal	1292	1535	11939	15680	2.26E-16	2.94E-15
			HSMMtube	normal	1236	1535	11545	15680	1.29E-11	1.68E-10
			NH-A	normal	1244	1535	11638	15680	1.43E-11	1.86E-10
			NHDF-Ad	normal	1239	1535	11380	15680	3.02E-15	3.92E-14
			NHEK	normal	1261	1535	11718	15680	1.32E-13	1.71E-12
			NHLF	normal	1251	1535	11406	15680	1.65E-17	2.14E-16
			K562	cancer	1082	1535	9977	15680	1.04E-09	1.35E-08
			HeLa-S3	cancer	1094	1535	10081	15680	4.52E-10	5.88E-09
			HepG2	cancer	1197	1535	11193	15680	2.99E-10	3.89E-09
			Dnd41	cancer	974	1535	8921	15680	1.63E-08	2.11E-07
		Intragenic CGIs	HUVEC	normal	29	235	441	3094	0.778663427	1
			CD14+	normal	33	235	533	3094	0.897225807	1
			HMEC	normal	35	235	421	3094	0.239519947	1
			HSMM	normal	42	235	536	3094	0.368616379	1
			HSMMtube	normal	38	235	529	3094	0.613033528	1
			NH-A	normal	29	235	489	3094	0.925426067	1
			NHDF-Ad	normal	37	235	449	3094	0.252882188	1
			NHEK	normal	37	235	494	3094	0.49441344	1
			NHLF	normal	30	235	432	3094	0.668744467	1
			K562	cancer	27	235	388	3094	0.649922962	1
			HeLa-S3	cancer	31	235	456	3094	0.721606582	1
			HepG2	cancer	44	235	461	3094	0.038337988	0.49839385
			Dnd41	cancer	14	235	230	3094	0.774964619	1
		3' CGIs	HUVEC	normal	44	143	473	1806	0.082899126	1

			CD14+	normal	47	143	533	1806	0.155772347	1
			HMEC	normal	46	143	437	1806	0.009110405	0.11843527
			HSMM	normal	47	143	521	1806	0.115614874	1
			HSMMtube	normal	43	143	464	1806	0.09035608	1
			NH-A	normal	44	143	467	1806	0.069095492	0.89824139
			NHDF-Ad	normal	45	143	471	1806	0.053754852	0.69881307
			NHEK	normal	54	143	522	1806	0.006561029	0.08529338
			NHLF	normal	45	143	453	1806	0.028522938	0.37079819
			K562	cancer	35	143	401	1806	0.214107716	1
			HeLa-S3	cancer	39	143	407	1806	0.066958041	0.87045453
			HepG2	cancer	41	143	435	1806	0.077186295	1
			Dnd41	cancer	28	143	252	1806	0.019055049	0.24771564
		Intergenic CGIs	HUVEC	normal	250	509	2626	6835	1.34E-07	1.75E-06
			CD14+	normal	252	509	2738	6835	2.94E-06	3.82E-05
			HMEC	normal	255	509	2681	6835	9.24E-08	1.20E-06
			HSMM	normal	283	509	3001	6835	1.45E-08	1.89E-07
			HSMMtube	normal	240	509	2670	6835	4.79E-05	0.00062262
			NH-A	normal	262	509	2813	6835	4.21E-07	5.48E-06
			NHDF-Ad	normal	243	509	2616	6835	2.49E-06	3.24E-05
			NHEK	normal	268	509	2876	6835	2.37E-07	3.09E-06
			NHLF	normal	236	509	2559	6835	7.67E-06	9.97E-05
			K562	cancer	193	509	2086	6835	8.73E-05	0.00113484
			HeLa-S3	cancer	169	509	2026	6835	0.031064262	0.4038354
			HepG2	cancer	240	509	2648	6835	2.48E-05	0.00032186
			Dnd41	cancer	147	509	1478	6835	2.43E-05	0.00031589
	5LSR	**5' CGIs**	HUVEC	normal	199	233	11399	15680	1.06E-06	1.37E-05
			CD14+	normal	193	233	11296	15680	3.89E-05	0.00050621
			HMEC	normal	195	233	11565	15680	0.000100394	1.31E-03
			HSMM	normal	208	233	11939	15680	8.45E-08	1.10E-06
			HSMMtube	normal	207	233	11545	15680	2.31E-09	3.00E-08
			NH-A	normal	204	233	11638	15680	1.50E-07	1.94E-06
			NHDF-Ad	normal	203	233	11380	15680	2.17E-08	2.82E-07
			NHEK	normal	199	233	11718	15680	2.31E-05	0.00029973
			NHLF	normal	197	233	11406	15680	5.72E-06	7.44E-05

			K562	cancer	170	233	9977	15680	0.000918968	0.01194659
			HeLa-S3	cancer	170	233	10081	15680	0.001822082	0.02368706
			HepG2	cancer	193	233	11193	15680	1.56E-05	0.00020322
			Dnd41	cancer	154	233	8921	15680	0.00156345	0.02032485
		Intragenic CGIs	HUVEC	normal	3	19	441	3094	0.281755207	1
			CD14+	normal	2	19	533	3094	0.660375345	1
			HMEC	normal	2	19	421	3094	0.489015704	1
			HSMM	normal	2	19	536	3094	0.664427789	1
			HSMMtube	normal	6	19	529	3094	0.031464488	0.40903835
			NH-A	normal	3	19	489	3094	0.352808918	1
			NHDF-Ad	normal	6	19	449	3094	0.01343878	0.17470413
			NHEK	normal	3	19	494	3094	0.360328784	1
			NHLF	normal	2	19	432	3094	0.507362718	1
			K562	cancer	1	19	388	3094	0.708891767	1
			HeLa-S3	cancer	1	19	456	3094	0.793780457	1
			HepG2	cancer	2	19	461	3094	0.554340034	1
			Dnd41	cancer	1	19	230	3094	0.41826206	1
		3' CGIs	HUVEC	normal	7	17	473	1806	0.051055748	0.66372472
			CD14+	normal	7	17	533	1806	0.095458345	1
			HMEC	normal	5	17	437	1806	0.20972921	1
			HSMM	normal	9	17	521	1806	0.009157676	0.11904978
			HSMMtube	normal	4	17	464	1806	0.452710588	1
			NH-A	normal	8	17	467	1806	0.0151013	0.19631689
			NHDF-Ad	normal	7	17	471	1806	0.049889915	0.64856889
			NHEK	normal	6	17	522	1806	0.194040971	1
			NHLF	normal	5	17	453	1806	0.236397344	1
			K562	cancer	2	17	401	1806	0.764175739	1
			HeLa-S3	cancer	2	17	407	1806	0.773970887	1
			HepG2	cancer	5	17	435	1806	0.206493324	1
			Dnd41	cancer	2	17	252	1806	0.430473217	1
		Intergenic CGIs	HUVEC	normal	35	65	2626	6835	0.00391671	0.05091723
			CD14+	normal	33	65	2738	6835	0.02985051	0.38805663
			HMEC	normal	30	65	2681	6835	0.101532922	1
			HSMM	normal	39	65	3001	6835	0.003048259	0.03962737

			HSMMtube	normal	32	65	2670	6835	0.035878191	0.46641649
			NH-A	normal	35	65	2813	6835	0.013964187	0.18153444
			NHDF-Ad	normal	32	65	2616	6835	0.026532101	0.34491731
			NHEK	normal	33	65	2876	6835	0.061037448	0.79348683
			NHLF	normal	26	65	2559	6835	0.286316305	1
			K562	cancer	25	65	2086	6835	0.065076454	0.8459939
			HeLa-S3	cancer	13	65	2026	6835	0.946000912	1
			HepG2	cancer	23	65	2648	6835	0.663529081	1
			Dnd41	cancer	11	65	1478	6835	0.776954598	1
	CE	**5'CGIs**	HUVEC	normal	7100	8782	11399	15680	1.89E-147	2.45E-146
			CD14+	normal	6987	8782	11296	15680	5.37E-124	6.99E-123
			HMEC	normal	7242	8782	11565	15680	2.17E-172	2.82E-171
			HSMM	normal	7472	8782	11939	15680	1.16E-193	1.50E-192
			HSMMtube	normal	7273	8782	11545	15680	2.98E-191	3.87E-190
			NH-A	normal	7282	8782	11638	15680	5.43E-174	7.05E-173
			NHDF-Ad	normal	7090	8782	11380	15680	2.60E-147	3.38E-146
			NHEK	normal	7338	8782	11718	15680	1.81E-181	2.35E-180
			NHLF	normal	7118	8782	11406	15680	2.10E-153	2.72E-152
			K562	cancer	6164	8782	9977	15680	4.99E-83	6.49E-82
			HeLa-S3	cancer	6293	8782	10081	15680	6.30E-105	8.18E-104
			HepG2	cancer	6986	8782	11193	15680	6.90E-144	8.97E-143
			Dnd41	cancer	5538	8782	8921	15680	9.48E-70	1.23E-68
		Intragenic CGIs	HUVEC	normal	223	1284	441	3094	1.33E-05	0.000173
			CD14+	normal	249	1284	533	3094	0.003209551	0.041724
			HMEC	normal	213	1284	421	3094	2.04E-05	0.000265
			HSMM	normal	251	1284	536	3094	0.002620	0.034066
			HSMMtube	normal	239	1284	529	3094	0.026758	0.347857
			NH-A	normal	233	1284	489	3094	0.001165	0.015150
			NHDF-Ad	normal	215	1284	449	3094	0.001315	0.017092
			NHEK	normal	247	1284	494	3094	1.28E-05	0.000167
			NHLF	normal	217	1284	432	3094	3.16E-05	0.000411
			K562	cancer	197	1284	388	3094	3.24E-05	0.000422
			HeLa-S3	cancer	215	1284	456	3094	0.003546	0.046092
			HepG2	cancer	219	1284	461	3094	0.002013	0.026163

			Dnd41	cancer	104	1284	230	3094	0.104339	1.000000
		3' CGIs	HUVEC	normal	265	911	473	1806	0.001972	0.025633
			CD14+	normal	295	911	533	1806	0.002973	0.038650
			HMEC	normal	247	911	437	1806	0.001453	0.018893
			HSMM	normal	283	911	521	1806	0.015761	0.204890
			HSMMtube	normal	249	911	464	1806	0.048052	0.624682
			NH-A	normal	250	911	467	1806	0.054213	0.704768
			NHDF-Ad	normal	247	911	471	1806	0.143967	1.000000
			NHEK	normal	290	911	522	1806	0.002365	0.030749
			NHLF	normal	242	911	453	1806	0.064310	0.836025
			K562	cancer	207	911	401	1806	0.277121	1.000000
			HeLa-S3	cancer	220	911	407	1806	0.043409	0.564319
			HepG2	cancer	227	911	435	1806	0.187133	1.000000
			Dnd41	cancer	129	911	252	1806	0.373133	1.000000
		Intergenic CGIs	HUVEC	normal	959	1686	2626	6835	3.57E-71	4.64E-70
			CD14+	normal	929	1686	2738	6835	1.40E-47	1.83E-46
			HMEC	normal	979	1686	2681	6835	1.05E-73	1.37E-72
			HSMM	normal	1090	1686	3001	6835	1.42E-87	1.85E-86
			HSMMtube	normal	1001	1686	2670	6835	2.46E-85	3.20E-84
			NH-A	normal	1069	1686	2813	6835	3.90E-101	5.07E-100
			NHDF-Ad	normal	944	1686	2616	6835	8.36E-66	1.09E-64
			NHEK	normal	1052	1686	2876	6835	2.19E-84	2.85E-83
			NHLF	normal	933	1686	2559	6835	1.71E-67	2.23E-66
			K562	cancer	725	1686	2086	6835	9.25E-37	1.20E-35
			HeLa-S3	cancer	731	1686	2026	6835	2.02E-44	2.62E-43
			HepG2	cancer	962	1686	2648	6835	5.75E-70	7.47E-69
			Dnd41	cancer	536	1686	1478	6835	2.78E-30	3.61E-29
H3K27ac	**HIR**	**5'CGIs**	HUVEC	normal	1019	1535	9179	15680	1.34E-11	1.74E-10
			CD14+	normal	964	1535	8858	15680	5.52E-08	7.18E-07
			HMEC	normal	986	1535	9036	15680	1.18E-08	1.53E-07
			HSMM	normal	962	1535	8902	15680	3.40E-07	4.42E-06
			HSMMtube	normal	931	1535	8631	15680	1.33E-06	1.72E-05
			NH-A	normal	996	1535	9121	15680	6.37E-09	8.28E-08
			NHDF-Ad	normal	1041	1535	9597	15680	6.52E-09	8.48E-08

			NHEK	normal	993	1535	9128	15680	2.05E-08	2.67E-07
			NHLF	normal	1028	1535	9402	15680	1.06E-09	1.37E-08
			K562	cancer	976	1535	8778	15680	8.33E-11	1.08E-09
			HeLa-S3	cancer	1013	1535	9136	15680	2.77E-11	3.61E-10
			HepG2	cancer	921	1535	8417	15680	6.61E-08	8.59E-07
			Dnd41	cancer	969	1535	8945	15680	1.48E-07	1.93E-06
		Intragenic CGIs	HUVEC	normal	21	235	340	3094	0.825340008	1
			CD14+	normal	38	235	583	3094	0.842215063	1
			HMEC	normal	10	235	225	3094	0.96411323	1
			HSMM	normal	19	235	357	3094	0.951734634	1
			HSMMtube	normal	23	235	411	3094	0.94253385	1
			NH-A	normal	14	235	322	3094	0.990235294	1
			NHDF-Ad	normal	31	235	519	3094	0.927750484	1
			NHEK	normal	11	235	280	3094	0.993154904	1
			NHLF	normal	22	235	398	3094	0.945532767	1
			K562	cancer	27	235	320	3094	0.234423785	1
			HeLa-S3	cancer	26	235	271	3094	0.081292371	1
			HepG2	cancer	19	235	229	3094	0.28507305	1
			Dnd41	cancer	62	235	662	3094	0.023544976	0.30608469
		3' CGIs	HUVEC	normal	29	143	306	1806	0.112093913	1
			CD14+	normal	40	143	435	1806	0.109799288	1
			HMEC	normal	24	143	239	1806	0.07918861	1
			HSMM	normal	29	143	296	1806	0.07962582	1
			HSMMtube	normal	32	143	311	1806	0.037860326	0.49218424
			NH-A	normal	31	143	268	1806	0.007863803	0.10222944
			NHDF-Ad	normal	32	143	403	1806	0.444516291	1
			NHEK	normal	25	143	299	1806	0.327921437	1
			NHLF	normal	35	143	353	1806	0.051481078	0.66925402
			K562	cancer	35	143	340	1806	0.030860906	0.40119177
			HeLa-S3	cancer	29	143	293	1806	0.071328609	0.92727191
			HepG2	cancer	28	143	249	1806	0.016271937	0.21153517
			Dnd41	cancer	40	143	420	1806	0.069747718	0.90672033
		Intergenic CGIs	HUVEC	normal	139	509	1488	6835	0.000874217	0.01136483
			CD14+	normal	129	509	1464	6835	0.011784574	0.15319946

			HMEC	normal	129	509	1352	6835	0.000586642	0.00762634
			HSMM	normal	124	509	1357	6835	0.004001398	0.05201818
			HSMMtube	normal	116	509	1354	6835	0.0366927	0.4770051
			NH-A	normal	145	509	1456	6835	2.63E-05	0.00034234
			NHDF-Ad	normal	145	509	1669	6835	0.012381231	0.160956
			NHEK	normal	123	509	1395	6835	0.013652571	0.17748343
			NHLF	normal	141	509	1511	6835	0.00083107	0.0108039
			K562	cancer	152	509	1472	6835	1.75E-06	2.28E-05
			HeLa-S3	cancer	133	509	1458	6835	0.003006324	0.03908221
			HepG2	cancer	104	509	1130	6835	0.006804828	0.08846277
			Dnd41	cancer	146	509	1591	6835	0.00137856	0.01792129
	5LSR	**5'CGIs**	HUVEC	normal	165	233	9179	15680	3.51E-05	0.00045644
			CD14+	normal	161	233	8858	15680	2.64E-05	0.00034381
			HMEC	normal	159	233	9036	15680	0.000315168	0.00409718
			HSMM	normal	163	233	8902	15680	1.14E-05	0.00014858
			HSMMtube	normal	161	233	8631	15680	3.67E-06	4.78E-05
			NH-A	normal	169	233	9121	15680	1.63E-06	2.13E-05
			NHDF-Ad	normal	178	233	9597	15680	2.49E-07	3.23E-06
			NHEK	normal	163	233	9128	15680	7.39E-05	0.00096055
			NHLF	normal	165	233	9402	15680	0.000200651	0.00260846
			K562	cancer	156	233	8778	15680	0.000225235	0.00292805
			HeLa-S3	cancer	157	233	9136	15680	0.001610201	0.02093262
			HepG2	cancer	154	233	8417	15680	4.10E-05	0.00053364
			Dnd41	cancer	150	233	8945	15680	0.009089806	0.11816747
		Intragenic CGIs	HUVEC	normal	2	19	340	3094	0.348214888	1
			CD14+	normal	5	19	583	3094	0.130980652	1
			HMEC	normal	1	19	225	3094	0.40713897	1
			HSMM	normal	2	19	357	3094	0.378242878	1
			HSMMtube	normal	1	19	411	3094	0.740205766	1
			NH-A	normal	2	19	322	3094	0.316466195	1
			NHDF-Ad	normal	7	19	519	3094	0.008014574	0.10418946
			NHEK	normal	1	19	280	3094	0.523921693	1
			NHLF	normal	0	19	398	3094	0.927517473	1
			K562	cancer	0	19	320	3094	0.8751593	1

			HeLa-S3	cancer	1	19	271	3094	0.505735219	1
			HepG2	cancer	0	19	229	3094	0.769030044	1
			Dnd41	cancer	10	19	662	3094	0.000550209	0.00715272
		3' CGIs	HUVEC	normal	0	17	306	1806	0.958059567	1
			CD14+	normal	3	17	435	1806	0.6148087	1
			HMEC	normal	1	17	239	1806	0.679660247	1
			HSMM	normal	2	17	296	1806	0.544954337	1
			HSMMtube	normal	3	17	311	1806	0.333942024	1
			NH-A	normal	3	17	268	1806	0.237498468	1
			NHDF-Ad	normal	3	17	403	1806	0.546503105	1
			NHEK	normal	0	17	299	1806	0.954585865	1
			NHLF	normal	2	17	353	1806	0.674970273	1
			K562	cancer	0	17	340	1806	0.971657448	1
			HeLa-S3	cancer	1	17	293	1806	0.789669019	1
			HepG2	cancer	0	17	249	1806	0.920669554	1
			Dnd41	cancer	2	17	420	1806	0.794178512	1
		Intergenic CGIs	HUVEC	normal	24	65	1488	6835	0.001599445	0.02079279
			CD14+	normal	16	65	1464	6835	0.213562803	1
			HMEC	normal	14	65	1352	6835	0.295865386	1
			HSMM	normal	14	65	1357	6835	0.301130624	1
			HSMMtube	normal	10	65	1354	6835	0.766900317	1
			NH-A	normal	14	65	1456	6835	0.410492354	1
			NHDF-Ad	normal	23	65	1669	6835	0.016222041	0.21088653
			NHEK	normal	13	65	1395	6835	0.459301223	1
			NHLF	normal	16	65	1511	6835	0.255970976	1
			K562	cancer	11	65	1472	6835	0.771856722	1
			HeLa-S3	cancer	10	65	1458	6835	0.847682447	1
			HepG2	cancer	11	65	1130	6835	0.387009036	1
			Dnd41	cancer	13	65	1591	6835	0.677252021	1
	CE	**5'CGIs**	HUVEC	normal	5730	8782	9179	15680	6.30E-83	8.19E-82
			CD14+	normal	5402	8782	8858	15680	7.92E-47	1.03E-45
			HMEC	normal	5652	8782	9036	15680	4.85E-83	6.31E-82
			HSMM	normal	5642	8782	8902	15680	2.17E-101	2.82E-100
			HSMMtube	normal	5495	8782	8631	15680	4.29E-102	5.58E-101

			NH-A	normal	5749	8782	9121	15680	1.86E-97	2.42E-96
			NHDF-Ad	normal	5979	8782	9597	15680	7.38E-89	9.60E-88
			NHEK	normal	5702	8782	9128	15680	6.54E-83	8.50E-82
			NHLF	normal	5849	8782	9402	15680	3.86E-82	5.02E-81
			K562	cancer	5457	8782	8778	15680	3.17E-69	4.12E-68
			HeLa-S3	cancer	5732	8782	9136	15680	4.33E-90	5.62E-89
			HepG2	cancer	5260	8782	8417	15680	5.68E-70	7.38E-69
			Dnd41	cancer	5496	8782	8945	15680	1.20E-56	1.56E-55
		Intragenic CGIs	HUVEC	normal	155	1284	340	3094	0.046864792	0.60924229
			CD14+	normal	256	1284	583	3094	0.087398419	1
			HMEC	normal	110	1284	225	3094	0.008335083	0.10835608
			HSMM	normal	153	1284	357	3094	0.270251272	1
			HSMMtube	normal	179	1284	411	3094	0.168309867	1
			NH-A	normal	141	1284	322	3094	0.173400386	1
			NHDF-Ad	normal	230	1284	519	3094	0.070208194	0.91270652
			NHEK	normal	137	1284	280	3094	0.00352142	0.04577846
			NHLF	normal	178	1284	398	3094	0.073442963	0.95475851
			K562	cancer	155	1284	320	3094	0.00339791	0.04417282
			HeLa-S3	cancer	142	1284	271	3094	5.93E-05	0.00077143
			HepG2	cancer	102	1284	229	3094	0.149102077	1
			Dnd41	cancer	270	1284	662	3094	0.646128785	1
		3' CGIs	HUVEC	normal	171	911	306	1806	0.0156651	0.2036463
			CD14+	normal	231	911	435	1806	0.091926603	1
			HMEC	normal	128	911	239	1806	0.134999356	1
			HSMM	normal	155	911	296	1806	0.215723479	1
			HSMMtube	normal	160	911	311	1806	0.32585844	1
			NH-A	normal	145	911	268	1806	0.086013542	1
			NHDF-Ad	normal	223	911	403	1806	0.011107431	0.1443966
			NHEK	normal	158	911	299	1806	0.165554665	1
			NHLF	normal	191	911	353	1806	0.055333839	0.71933991
			K562	cancer	173	911	340	1806	0.405200056	1
			HeLa-S3	cancer	162	911	293	1806	0.030181149	0.39235494
			HepG2	cancer	132	911	249	1806	0.173234529	1
			Dnd41	cancer	222	911	420	1806	0.117934346	1

		Intergenic CGIs	HUVEC	normal	579	1686	1488	6835	1.45E-44	1.88E-43
			CD14+	normal	515	1686	1464	6835	4.77E-25	6.20E-24
			HMEC	normal	523	1686	1352	6835	1.96E-38	2.55E-37
			HSMM	normal	529	1686	1357	6835	3.86E-40	5.02E-39
			HSMMtube	normal	542	1686	1354	6835	1.32E-45	1.72E-44
			NH-A	normal	579	1686	1456	6835	2.50E-48	3.26E-47
			NHDF-Ad	normal	614	1686	1669	6835	2.98E-38	3.87E-37
			NHEK	normal	514	1686	1395	6835	7.31E-31	9.51E-30
			NHLF	normal	576	1686	1511	6835	7.18E-41	9.34E-40
			K562	cancer	514	1686	1472	6835	4.17E-24	5.43E-23
			HeLa-S3	cancer	563	1686	1458	6835	8.78E-42	1.14E-40
			HepG2	cancer	463	1686	1130	6835	4.99E-41	6.48E-40
			Dnd41	cancer	563	1686	1591	6835	9.92E-29	1.29E-27
H3K36me3	**HIR**	**5'CGIs**	HUVEC	normal	669	1535	6558	15680	0.067210303	0.87373394
			CD14+	normal	519	1535	4950	15680	0.022188677	0.28845281
			HMEC	normal	660	1535	6669	15680	0.338733884	1
			HSMM	normal	848	1535	8023	15680	0.000343315	0.00446309
			HSMMtube	normal	677	1535	6641	15680	0.068423821	0.88950968
			NH-A	normal	613	1535	5938	15680	0.037534781	0.48795215
			NHDF-Ad	normal	748	1535	7065	15680	0.001082305	0.01406997
			NHEK	normal	769	1535	7533	15680	0.042381947	0.55096531
			NHLF	normal	769	1535	7437	15680	0.012882143	0.16746786
			K562	cancer	784	1535	6949	15680	9.51E-09	1.24E-07
			HeLa-S3	cancer	735	1535	6665	15680	3.46E-06	4.50E-05
			HepG2	cancer	668	1535	6288	15680	1.91E-03	0.02482775
			Dnd41	cancer	818	1535	7612	15680	4.04E-05	0.0005256
		Intragenic CGIs	HUVEC	normal	86	235	1280	3094	0.930798207	1
			CD14+	normal	64	235	1188	3094	0.999875172	1
			HMEC	normal	85	235	1232	3094	0.868747145	1
			HSMM	normal	100	235	1436	3094	0.878324946	1
			HSMMtube	normal	107	235	1577	3094	0.952244821	1
			NH-A	normal	76	235	1056	3094	0.700477153	1
			NHDF-Ad	normal	86	235	1250	3094	0.878795401	1
			NHEK	normal	90	235	1338	3094	0.936758049	1

			NHLF	normal	91	235	1495	3094	0.998680813	1
			K562	cancer	62	235	1044	3094	0.992814645	1
			HeLa-S3	cancer	92	235	1119	3094	0.144614695	1
			HepG2	cancer	91	235	1313	3094	0.870864711	1
			Dnd41	cancer	84	235	1301	3094	0.976081611	1
		3' CGIs	HUVEC	normal	77	143	770	1806	0.001885221	0.02450787
			CD14+	normal	63	143	662	1806	0.023435846	0.304666
			HMEC	normal	77	143	736	1806	0.000363765	0.00472894
			HSMM	normal	75	143	854	1806	0.084591729	1
			HSMMtube	normal	80	143	861	1806	0.015766181	0.20496035
			NH-A	normal	65	143	608	1806	0.000841661	0.01094159
			NHDF-Ad	normal	73	143	754	1806	0.007669583	0.09970458
			NHEK	normal	71	143	828	1806	0.149517063	1
			NHLF	normal	77	143	876	1806	0.077955821	1
			K562	cancer	61	143	630	1806	0.017800492	0.23140639
			HeLa-S3	cancer	64	143	643	1806	0.007287572	0.09473843
			HepG2	cancer	72	143	679	1806	0.000441824	0.00574371
			Dnd41	cancer	71	143	771	1806	0.033247076	0.43221198
		Intergenic CGIs	HUVEC	normal	59	509	678	6835	0.084698387	1
			CD14+	normal	56	509	605	6835	0.03471042	0.45123546
			HMEC	normal	67	509	791	6835	0.109272568	1
			HSMM	normal	100	509	1177	6835	0.060238262	0.7830974
			HSMMtube	normal	83	509	873	6835	0.006546956	0.08511043
			NH-A	normal	58	509	600	6835	0.014536252	0.18897127
			NHDF-Ad	normal	72	509	818	6835	0.052539249	0.68301023
			NHEK	normal	79	509	898	6835	0.044882391	0.58347108
			NHLF	normal	86	509	1026	6835	0.097838329	1
			K562	cancer	93	509	846	6835	2.50E-05	0.00032537
			HeLa-S3	cancer	73	509	814	6835	0.035945814	0.46729558
			HepG2	cancer	68	509	737	6835	0.024010632	0.31213821
			Dnd41	cancer	85	509	957	6835	0.031571295	0.41042684
	5LSR	**5'CGIs**	HUVEC	normal	104	233	6558	15680	0.172649457	1
			CD14+	normal	84	233	4950	15680	0.061356652	0.79763648
			HMEC	normal	106	233	6669	15680	0.161545556	1

			HSMM	normal	129	233	8023	15680	0.087179711	1
			HSMMtube	normal	101	233	6641	15680	0.35234683	1
			NH-A	normal	101	233	5938	15680	0.036374095	0.47286323
			NHDF-Ad	normal	127	233	7065	15680	0.001448573	0.01883145
			NHEK	normal	120	233	7533	15680	0.129029023	1
			NHLF	normal	122	233	7437	15680	0.056614077	0.73598301
			K562	cancer	133	233	6949	15680	3.14E-05	0.00040782
			HeLa-S3	cancer	138	233	6665	15680	8.41E-08	1.09E-06
			HepG2	cancer	112	233	6288	15680	0.005409569	0.0703244
			Dnd41	cancer	127	233	7612	15680	0.02870035	0.37310455
		Intragenic CGIs	HUVEC	normal	13	19	1280	3094	0.004338901	0.05640571
			CD14+	normal	11	19	1188	3094	0.024852816	0.32308661
			HMEC	normal	9	19	1232	3094	0.180989517	1
			HSMM	normal	13	19	1436	3094	0.014809112	0.19251846
			HSMMtube	normal	13	19	1577	3094	0.037843282	0.49196267
			NH-A	normal	12	19	1056	3094	0.002324823	0.03022271
			NHDF-Ad	normal	13	19	1250	3094	0.003339216	0.04340981
			NHEK	normal	12	19	1338	3094	0.023605472	0.30687114
			NHLF	normal	13	19	1495	3094	0.022335542	0.29036204
			K562	cancer	8	19	1044	3094	0.15460513	1
			HeLa-S3	cancer	12	19	1119	3094	0.004203321	0.05464317
			HepG2	cancer	13	19	1313	3094	0.005729294	0.07448082
			Dnd41	cancer	10	19	1301	3094	0.121440184	1
		3' CGIs	HUVEC	normal	4	17	770	1806	0.914714111	1
			CD14+	normal	4	17	662	1806	0.807536039	1
			HMEC	normal	6	17	736	1806	0.577857379	1
			HSMM	normal	7	17	854	1806	0.601613746	1
			HSMMtube	normal	4	17	861	1806	0.962937018	1
			NH-A	normal	3	17	608	1806	0.876651482	1
			NHDF-Ad	normal	3	17	754	1806	0.966503609	1
			NHEK	normal	6	17	828	1806	0.734532897	1
			NHLF	normal	5	17	876	1806	0.910791987	1
			K562	cancer	7	17	630	1806	0.208742787	1
			HeLa-S3	cancer	6	17	643	1806	0.401362869	1

			HepG2	cancer	10	17	679	1806	0.020970127	0.27261166
			Dnd41	cancer	7	17	771	1806	0.447979481	1
		Intergenic CGIs	HUVEC	normal	13	65	678	6835	0.003946679	0.05130682
			CD14+	normal	5	65	605	6835	0.520344185	1
			HMEC	normal	5	65	791	6835	0.779037215	1
			HSMM	normal	16	65	1177	6835	0.045130092	0.5866912
			HSMMtube	normal	6	65	873	6835	0.741106363	1
			NH-A	normal	4	65	600	6835	0.68637612	1
			NHDF-Ad	normal	8	65	818	6835	0.374187104	1
			NHEK	normal	8	65	898	6835	0.487839491	1
			NHLF	normal	7	65	1026	6835	0.779964351	1
			K562	cancer	4	65	846	6835	0.918251224	1
			HeLa-S3	cancer	5	65	814	6835	0.802508545	1
			HepG2	cancer	4	65	737	6835	0.844410487	1
			Dnd41	cancer	8	65	957	6835	0.569155962	1
	CE	**5'CGIs**	HUVEC	normal	4076	8782	6558	15680	4.75E-40	6.18E-39
			CD14+	normal	3025	8782	4950	15680	7.66E-19	9.96E-18
			HMEC	normal	4138	8782	6669	15680	7.80E-40	1.01E-38
			HSMM	normal	4994	8782	8023	15680	6.16E-59	8.01E-58
			HSMMtube	normal	4104	8782	6641	15680	1.65E-36	2.15E-35
			NH-A	normal	3729	8782	5938	15680	1.81E-41	2.35E-40
			NHDF-Ad	normal	4479	8782	7065	15680	1.06E-64	1.38E-63
			NHEK	normal	4684	8782	7533	15680	3.02E-51	3.93E-50
			NHLF	normal	4653	8782	7437	15680	2.91E-56	3.79E-55
			K562	cancer	4335	8782	6949	15680	2.69E-47	3.49E-46
			HeLa-S3	cancer	4211	8782	6665	15680	2.55E-55	3.31E-54
			HepG2	cancer	3903	8782	6288	15680	1.65E-36	2.15E-35
			Dnd41	cancer	4723	8782	7612	15680	4.25E-50	5.53E-49
		Intragenic CGIs	HUVEC	normal	590	1284	1280	3094	5.68E-06	7.38E-05
			CD14+	normal	505	1284	1188	3094	0.17447696	1
			HMEC	normal	590	1284	1232	3094	1.85E-09	2.40E-08
			HSMM	normal	648	1284	1436	3094	6.02E-05	0.00078251
			HSMMtube	normal	704	1284	1577	3094	0.000128638	0.0016723
			NH-A	normal	494	1284	1056	3094	7.78E-06	0.00010115

			NHDF-Ad	normal	570	1284	1250	3094	6.04E-05	0.00078514
			NHEK	normal	612	1284	1338	3094	1.26E-05	0.00016395
			NHLF	normal	677	1284	1495	3094	1.53E-05	0.00019919
			K562	cancer	473	1284	1044	3094	0.000966049	0.01255863
			HeLa-S3	cancer	529	1284	1119	3094	4.00E-07	5.20E-06
			HepG2	cancer	550	1284	1313	3094	0.339310805	1
			Dnd41	cancer	565	1284	1301	3094	0.02933107	0.38130392
		3' CGIs	HUVEC	normal	420	911	770	1806	0.001124399	0.01461718
			CD14+	normal	353	911	662	1806	0.02796486	0.36354318
			HMEC	normal	408	911	736	1806	0.000178937	0.00232618
			HSMM	normal	469	911	854	1806	0.000130139	0.0016918
			HSMMtube	normal	475	911	861	1806	5.15E-05	0.000669
			NH-A	normal	341	911	608	1806	0.000260584	0.00338759
			NHDF-Ad	normal	418	911	754	1806	0.000133994	0.00174192
			NHEK	normal	463	911	828	1806	7.36E-06	9.56E-05
			NHLF	normal	485	911	876	1806	1.97E-05	0.00025634
			K562	cancer	343	911	630	1806	0.005544192	0.07207449
			HeLa-S3	cancer	357	911	643	1806	0.00055535	0.00721955
			HepG2	cancer	377	911	679	1806	0.00033387	0.00434031
			Dnd41	cancer	411	911	771	1806	0.015801376	0.20541789
		Intergenic CGIs	HUVEC	normal	303	1686	678	6835	6.49E-34	8.43E-33
			CD14+	normal	227	1686	605	6835	5.94E-14	7.72E-13
			HMEC	normal	329	1686	791	6835	1.59E-29	2.06E-28
			HSMM	normal	467	1686	1177	6835	5.80E-37	7.53E-36
			HSMMtube	normal	347	1686	873	6835	1.08E-26	1.41E-25
			NH-A	normal	266	1686	600	6835	6.49E-29	8.43E-28
			NHDF-Ad	normal	354	1686	818	6835	2.54E-36	3.30E-35
			NHEK	normal	373	1686	898	6835	1.14E-33	1.48E-32
			NHLF	normal	428	1686	1026	6835	7.80E-40	1.01E-38
			K562	cancer	341	1686	846	6835	1.53E-27	1.99E-26
			HeLa-S3	cancer	336	1686	814	6835	1.96E-29	2.54E-28
			HepG2	cancer	302	1686	737	6835	1.19E-25	1.55E-24
			Dnd41	cancer	364	1686	957	6835	8.89E-24	1.16E-22

Supplementary table 3.4: Raw data used to calculate CGIs enriched with H3K4me3, H3K27ac and H3K36me3 in different cell lines according to the CGIs evolutionary model.

Histone mark	Signature of selective pressure	CGIs according to evolutionary model	Cell lines	Category	No.of CGIs with peaks under selective pressure (k)	No.of CGIs under selective pressure (n)	No. of CGIs with peaks (M)	Total no. of CGIs (N)	Hyper-geometric P-value	bonferroni P-value
H3K4me3	**HIR**	**Hypodeaminated CGIs**	HUVEC	normal	731	907	6957	9091	0.000834	0.01083714
			CD14+	normal	720	907	6935	9091	0.008619	0.11204396
			HMEC	normal	735	907	7063	9091	0.004263	0.05542167
			HSMM	normal	766	907	7317	9091	0.000487	0.0063369
			HSMMtube	normal	717	907	7007	9091	0.061529	0.79987202
			NH-A	normal	737	907	7158	9091	0.021772	0.28303785
			NHDF-Ad	normal	718	907	6952	9091	0.018996	0.24695346
			NHEK	normal	756	907	7254	9091	0.0018	0.02339949
			NHLF	normal	715	907	6910	9091	0.015368	0.19978297
			K562	cancer	632	907	6002	9091	0.00609	0.07916973
			HeLaS3	cancer	621	907	5934	9091	0.014696	0.19104443
			HepG2	cancer	703	907	6754	9091	0.008171	0.10622555
			Dnd41	cancer	551	907	5169	9091	0.005585	0.07261111
		BGC CGIs	HUVEC	normal	129	345	1204	4782	6.40E-08	8.32E-07
			CD14+	normal	131	345	1341	4782	1.27E-05	1.65E-04
			HMEC	normal	122	345	1162	4782	6.10E-07	7.92E-06
			HSMM	normal	135	345	1370	4782	5.10E-06	6.63E-05
			HSMMtube	normal	125	345	1294	4782	4.09E-05	5.32E-04
			NH-A	normal	124	345	1203	4782	1.36E-06	1.77E-05
			NHDF-Ad	normal	134	345	1220	4782	5.28E-09	6.87E-08
			NHEK	normal	128	345	1259	4782	1.83E-06	2.38E-05
			NHLF	normal	117	345	1174	4782	1.84E-05	2.40E-04
			K562	cancer	113	345	1112	4782	1.07E-05	1.39E-04
			HeLaS3	cancer	126	345	1286	4782	1.78E-05	2.32E-04
			HepG2	cancer	128	345	1302	4782	1.22E-05	1.58E-04
			Dnd41	cancer	103	345	853	4782	4.35E-09	5.65E-08
	5LSR	**Hypodeaminated CGIs**	HUVEC	normal	108	138	6957	9091	0.283012	1
			CD14+	normal	102	138	6935	9091	0.715888	1
			HMEC	normal	109	138	7063	9091	0.324192	1
			HSMM	normal	117	138	7317	9091	0.078514	1

			HSMMtube	normal	114	138	7007	9091	0.044844	0.58296702
			NH-A	normal	109	138	7158	9091	0.437416	1
			NHDF-Ad	normal	113	138	6952	9091	0.050077	0.65100448
			NHEK	normal	112	138	7254	9091	0.310684	1
			NHLF	normal	106	138	6910	9091	0.379269	1
			K562	cancer	83	138	6002	9091	0.914782	1
			HeLaS3	cancer	83	138	5934	9091	0.881381	1
			HepG2	cancer	97	138	6754	9091	0.838183	1
			Dnd41	cancer	75	138	5169	9091	0.697099	1
		BGC CGIs	HUVEC	normal	9	28	1204	4782	0.142858	1
			CD14+	normal	6	28	1341	4782	0.708068	1
			HMEC	normal	8	28	1162	4782	0.221869	1
			HSMM	normal	9	28	1370	4782	0.262014	1
			HSMMtube	normal	10	28	1294	4782	0.10857	1
			NH-A	normal	7	28	1203	4782	0.407707	1
			NHDF-Ad	normal	10	28	1220	4782	0.07636	0.99267721
			NHEK	normal	8	28	1259	4782	0.304875	1
			NHLF	normal	8	28	1174	4782	0.231545	1
			K562	cancer	6	28	1112	4782	0.485857	1
			HeLaS3	cancer	8	28	1286	4782	0.329635	1
			HepG2	cancer	8	28	1302	4782	0.344566	1
			Dnd41	cancer	6	28	853	4782	0.220868	1
	CE	**Hypodeaminated CGIs**	HUVEC	normal	3913	4881	6957	9091	4.90E-19	6.38E-18
			CD14+	normal	3851	4881	6935	9091	1.26E-10	1.63E-09
			HMEC	normal	3974	4881	7063	9091	1.73E-20	2.25E-19
			HSMM	normal	4119	4881	7317	9091	2.10E-24	2.73E-23
			HSMMtube	normal	3986	4881	7007	9091	1.59E-29	2.07E-28
			NH-A	normal	4012	4881	7158	9091	1.73E-18	2.25E-17
			NHDF-Ad	normal	3877	4881	6952	9091	3.51E-13	4.56E-12
			NHEK	normal	4055	4881	7254	9091	2.00E-17	2.59E-16
			NHLF	normal	3872	4881	6910	9091	6.47E-16	8.41E-15
			K562	cancer	3373	4881	6002	9091	1.03E-11	1.34E-10
			HeLaS3	cancer	3418	4881	5934	9091	4.88E-25	6.35E-24
			HepG2	cancer	3820	4881	6754	9091	4.78E-21	6.22E-20

			Dnd41	cancer	2994	4881	5169	9091	6.24E-21	8.11E-20
		BGC CGIs	HUVEC	normal	438	1338	1204	4782	5.89E-14	7.66E-13
			CD14+	normal	454	1338	1341	4782	9.32E-09	1.21E-07
			HMEC	normal	418	1338	1162	4782	2.59E-12	3.37E-11
			HSMM	normal	472	1338	1370	4782	1.69E-10	2.20E-09
			HSMMtube	normal	443	1338	1294	4782	2.69E-09	3.50E-08
			NH-A	normal	428	1338	1203	4782	9.06E-12	1.18E-10
			NHDF-Ad	normal	418	1338	1220	4782	9.05E-09	1.18E-07
			NHEK	normal	446	1338	1259	4782	5.43E-12	7.06E-11
			NHLF	normal	417	1338	1174	4782	2.66E-11	3.46E-10
			K562	cancer	412	1338	1112	4782	1.64E-14	2.13E-13
			HeLaS3	cancer	448	1338	1286	4782	1.02E-10	1.32E-09
			HepG2	cancer	461	1338	1302	4782	1.99E-12	2.59E-11
			Dnd41	cancer	320	1338	853	4782	8.18E-12	1.06E-10
H3k27ac	**HIR**	**Hypodeaminated CGIs**	HUVEC	normal	562	907	5346	9091	0.018869	0.24530313
			CD14+	normal	534	907	5121	9091	0.047821	0.62167341
			HMEC	normal	549	907	5298	9091	0.068502	0.89053238
			HSMM	normal	534	907	5148	9091	0.069877	0.90840636
			HSMMtube	normal	504	907	4982	9091	0.300404	1
			NH-A	normal	556	907	5288	9091	0.019826	0.25773606
			NHDF-Ad	normal	575	907	5632	9091	0.16347	1
			NHEK	normal	548	907	5335	9091	0.124198	1
			NHLF	normal	569	907	5446	9091	0.030535	0.39695404
			K562	cancer	557	907	5159	9091	0.001202	0.01562905
			HeLaS3	cancer	561	907	5304	9091	0.01066	0.1385841
			HepG2	cancer	522	907	4831	9091	0.002206	0.02867685
			Dnd41	cancer	558	907	5186	9091	0.001766	0.02295731
		BGC CGIs	HUVEC	normal	117	345	985	4782	6.59E-10	8.57E-09
			CD14+	normal	121	345	1289	4782	0.000227	2.95E-03
			HMEC	normal	98	345	801	4782	5.81E-09	7.55E-08
			HSMM	normal	108	345	933	4782	1.91E-08	2.49E-07
			HSMMtube	normal	111	345	995	4782	9.53E-08	1.24E-06
			NH-A	normal	98	345	870	4782	4.56E-07	5.93E-06
			NHDF-Ad	normal	121	345	1165	4782	1.29E-06	1.68E-05

			NHEK	normal	99	345	892	4782	8.17E-07	1.06E-05
			NHLF	normal	109	345	1005	4782	6.14E-07	7.98E-06
			K562	cancer	116	345	971	4782	5.58E-10	7.25E-09
			HeLaS3	cancer	106	345	942	4782	1.33E-07	1.73E-06
			HepG2	cancer	91	345	824	4782	3.30E-06	4.30E-05
			Dnd41	cancer	141	345	1377	4782	2.21E-07	2.87E-06
	5LSR	**Hypodeaminated CGIs**	HUVEC	normal	88	138	5346	9091	0.099486	1
			CD14+	normal	84	138	5121	9091	0.120714	1
			HMEC	normal	81	138	5298	9091	0.427491	1
			HSMM	normal	82	138	5148	9091	0.226076	1
			HSMMtube	normal	81	138	4982	9091	0.155651	1
			NH-A	normal	85	138	5288	9091	0.181833	1
			NHDF-Ad	normal	98	138	5632	9091	0.009766	0.12695304
			NHEK	normal	80	138	5335	9091	0.535547	1
			NHLF	normal	87	138	5446	9091	0.199408	1
			K562	cancer	74	138	5159	9091	0.746042	1
			HeLaS3	cancer	77	138	5304	9091	0.701113	1
			HepG2	cancer	70	138	4831	9091	0.687313	1
			Dnd41	cancer	75	138	5186	9091	0.712584	1
		BGC CGIs	HUVEC	normal	8	28	985	4782	0.103776	1
			CD14+	normal	11	28	1289	4782	0.050066	0.65085246
			HMEC	normal	7	28	801	4782	0.082873	1
			HSMM	normal	9	28	933	4782	0.033046	0.42959758
			HSMMtube	normal	9	28	995	4782	0.049244	0.64017842
			NH-A	normal	7	28	870	4782	0.121137	1
			NHDF-Ad	normal	12	28	1165	4782	0.008794	0.11431614
			NHEK	normal	7	28	892	4782	0.135173	1
			NHLF	normal	6	28	1005	4782	0.37191	1
			K562	cancer	6	28	971	4782	0.336307	1
			HeLaS3	cancer	7	28	942	4782	0.17026	1
			HepG2	cancer	6	28	824	4782	0.195344	1
			Dnd41	cancer	13	28	1377	4782	0.014165	0.18414148
	CE	**Hypodeaminated CGIs**	HUVEC	normal	3092	4881	5346	9091	1.08E-21	1.41E-20
			CD14+	normal	2916	4881	5121	9091	7.06E-13	9.18E-12

			HMEC	normal	3069	4881	5298	9091	4.09E-22	5.31E-21
			HSMM	normal	3040	4881	5148	9091	3.87E-32	5.03E-31
			HSMMtube	normal	2953	4881	4982	9091	2.34E-32	3.04E-31
			NH-A	normal	3085	4881	5288	9091	4.02E-26	5.23E-25
			NHDF-Ad	normal	3212	4881	5632	9091	1.53E-16	1.99E-15
			NHEK	normal	3076	4881	5335	9091	6.46E-20	8.40E-19
			NHLF	normal	3124	4881	5446	9091	3.82E-18	4.97E-17
			K562	cancer	2966	4881	5159	9091	3.47E-17	4.51E-16
			HeLaS3	cancer	3109	4881	5304	9091	2.81E-29	3.66E-28
			HepG2	cancer	2830	4881	4831	9091	8.99E-24	1.17E-22
			Dnd41	cancer	2994	4881	5186	9091	2.15E-19	2.79E-18
		BGC CGIs	HUVEC	normal	368	1338	985	4782	2.09E-13	2.72E-12
			CD14+	normal	456	1338	1289	4782	3.41E-12	4.43E-11
			HMEC	normal	314	1338	801	4782	1.51E-14	1.96E-13
			HSMM	normal	345	1338	933	4782	8.67E-12	1.13E-10
			HSMMtube	normal	366	1338	995	4782	3.50E-12	4.55E-11
			NH-A	normal	330	1338	870	4782	5.91E-13	7.69E-12
			NHDF-Ad	normal	418	1338	1165	4782	4.14E-12	5.38E-11
			NHEK	normal	332	1338	892	4782	9.48E-12	1.23E-10
			NHLF	normal	366	1338	1005	4782	1.79E-11	2.32E-10
			K562	cancer	371	1338	971	4782	2.69E-15	3.49E-14
			HeLaS3	cancer	365	1338	942	4782	3.68E-16	4.79E-15
			HepG2	cancer	331	1338	824	4782	2.69E-17	3.50E-16
			Dnd41	cancer	469	1338	1377	4782	1.55E-09	2.02E-08
H3K36me3	**HIR**	**Hypodeaminated CGIs**	HUVEC	normal	370	907	3612	9091	0.234017	1
			CD14+	normal	299	907	2750	9091	0.028395	0.36912887
			HMEC	normal	363	907	3702	9091	0.66087	1
			HSMM	normal	465	907	4497	9091	0.11927	1
			HSMMtube	normal	361	907	3630	9091	0.518084	1
			NH-A	normal	338	907	3260	9091	0.166704	1
			NHDF-Ad	normal	397	907	3833	9091	0.142562	1
			NHEK	normal	417	907	4154	9091	0.414611	1
			NHLF	normal	403	907	4029	9091	0.456648	1
			K562	cancer	426	907	3867	9091	0.00204	0.0265251

			HeLaS3	cancer	402	907	3736	9091	0.017354	0.22560689
			HepG2	cancer	374	907	3435	9091	0.011137	0.14477735
			Dnd41	cancer	460	907	4143	9091	0.00047	0.006111
		BGC CGIs	HUVEC	normal	125	345	1682	4782	0.31221	1
			CD14+	normal	115	345	1591	4782	0.463909	1
			HMEC	normal	129	345	1630	4782	0.080941	1
			HSMM	normal	153	345	1943	4782	0.065207	0.84769658
			HSMMtube	normal	140	345	1997	4782	0.656461	1
			NH-A	normal	103	345	1329	4782	0.170709	1
			NHDF-Ad	normal	132	345	1622	4782	0.034679	0.45082273
			NHEK	normal	148	345	1824	4782	0.026442	0.34374195
			NHLF	normal	140	345	1996	4782	0.653451	1
			K562	cancer	124	345	1489	4782	0.020572	0.26743116
			HeLaS3	cancer	124	345	1504	4782	0.028093	0.36521405
			HepG2	cancer	128	345	1686	4782	0.210546	1
			Dnd41	cancer	143	345	1915	4782	0.270525	1
	5LSR	**Hypodeaminated CGIs**	HUVEC	normal	54	138	3612	9091	0.520706	1
			CD14+	normal	43	138	2750	9091	0.367571	1
			HMEC	normal	45	138	3702	9091	0.970201	1
			HSMM	normal	75	138	4497	9091	0.107167	1
			HSMMtube	normal	56	138	3630	9091	0.401315	1
			NH-A	normal	48	138	3260	9091	0.566845	1
			NHDF-Ad	normal	66	138	3833	9091	0.074836	0.97287081
			NHEK	normal	59	138	4154	9091	0.729331	1
			NHLF	normal	63	138	4029	9091	0.342154	1
			K562	cancer	60	138	3867	9091	0.376061	1
			HeLaS3	cancer	66	138	3736	9091	0.044661	0.58058869
			HepG2	cancer	55	138	3435	9091	0.27483	1
			Dnd41	cancer	62	138	4143	9091	0.525789	1
		BGC CGIs	HUVEC	normal	14	28	1682	4782	0.034768	0.45198332
			CD14+	normal	12	28	1591	4782	0.101961	1
			HMEC	normal	14	28	1630	4782	0.026147	0.33991207
			HSMM	normal	18	28	1943	4782	0.003222	0.04187976
			HSMMtube	normal	15	28	1997	4782	0.072655	0.94451966

			NH-A	normal	13	28	1329	4782	0.010215	0.13279852
			NHDF-Ad	normal	16	28	1622	4782	0.003253	0.04229248
			NHEK	normal	15	28	1824	4782	0.031639	0.41131271
			NHLF	normal	15	28	1996	4782	0.072339	0.94041122
			K562	cancer	14	28	1489	4782	0.011017	0.14322697
			HeLaS3	cancer	14	28	1504	4782	0.012161	0.15808916
			HepG2	cancer	15	28	1686	4782	0.014433	0.18762763
			Dnd41	cancer	14	28	1915	4782	0.102705	1
	CE	**Hypodeaminated CGIs**	HUVEC	normal	2187	4881	3612	9091	5.42E-27	7.04E-26
			CD14+	normal	1623	4881	2750	9091	7.57E-12	9.84E-11
			HMEC	normal	2230	4881	3702	9091	1.00E-25	1.30E-24
			HSMM	normal	2665	4881	4497	9091	1.99E-26	2.59E-25
			HSMMtube	normal	2201	4881	3630	9091	7.82E-28	1.02E-26
			NH-A	normal	1986	4881	3260	9091	1.39E-25	1.81E-24
			NHDF-Ad	normal	2369	4881	3833	9091	1.03E-40	1.34E-39
			NHEK	normal	2502	4881	4154	9091	5.62E-31	7.31E-30
			NHLF	normal	2429	4881	4029	9091	6.74E-30	8.76E-29
			K562	cancer	2312	4881	3867	9091	3.72E-24	4.83E-23
			HeLaS3	cancer	2256	4881	3736	9091	3.37E-27	4.38E-26
			HepG2	cancer	2094	4881	3435	9091	6.73E-28	8.74E-27
			Dnd41	cancer	2525	4881	4143	9091	1.69E-37	2.19E-36
		BGC CGIs	HUVEC	normal	674	1338	1682	4782	2.55E-42	3.31E-41
			CD14+	normal	602	1338	1591	4782	1.22E-26	1.58E-25
			HMEC	normal	646	1338	1630	4782	1.23E-37	1.60E-36
			HSMM	normal	729	1338	1943	4782	4.06E-34	5.28E-33
			HSMMtube	normal	750	1338	1997	4782	5.40E-36	7.02E-35
			NH-A	normal	541	1338	1329	4782	3.20E-33	4.16E-32
			NHDF-Ad	normal	639	1338	1622	4782	5.90E-36	7.67E-35
			NHEK	normal	720	1338	1824	4782	1.04E-43	1.35E-42
			NHLF	normal	766	1338	1996	4782	5.09E-42	6.61E-41
			K562	cancer	566	1338	1489	4782	4.65E-25	6.04E-24
			HeLaS3	cancer	586	1338	1504	4782	5.36E-30	6.96E-29
			HepG2	cancer	618	1338	1686	4782	5.99E-23	7.78E-22
			Dnd41	cancer	716	1338	1915	4782	1.84E-32	2.39E-31

List of abbreviations

DNMTs : DNA methyltransferases

CGIs : CpG islands

H3K4me3 : Histone H3 lysine 4 trimethylation

H3K27ac : Histone H3 lysine 27acetylation

H3K36me3 : Histone H3 lysine 36 trimethylation

NuRD : Nucleosome remodelling and deacetylation complex

SETD2 : SET domain containing protein 2 (Histone-lysine N-methyltransferase)

JHDM3A : Jumonji C (JmjC)-domain-containing Histone Demethylase 3A

Cfp1 : CxxC finger protein 1

MBDs : Methyl CpG Binding Domain Proteins

HDAC : Histone Deacetylase

PRC1 : Polycomb Repressive Complex 1

PRC2 : Polycomb Repressive Complex 2

iHS : Integrated Haplotype Score

HIR : High iHS Regions

S score : Selective sweep scan score

5LSR : 5% Lowest S Regions

CE : Conserved Elements

CT : Cancer Transformed

EBV : Epstein Barr Virus transformed

NU : Normal Untransformed

TSS : Transcription Start Sites

UCSC GB : University of California Santa Cruz Genome Browser

ENCODE : Encyclopedia of DNA Elements

PS : Peaks Signal

BGC : Biased Gene Conversion

Acknowledgement

First of all, I would like to express my sincere acknowledgement and deepest appreciation to my tutor Prof. Sergio Cocozza for his kindness, patience, advice, supervision and tremendous support over the past three years of my PhD.

I would also like to sincerely acknowledge my co-tutor Prof. Gennaro Miele for his valuable suggestions, critical discussion and continious guidance to carry out this project.

I would like to give special thanks to Dr. Antonella Monticelli who is a researcher at IEOS, CNR, not only for her suggestions, criticism and enormous help for this project but also for her care and sensibility about me that made me feel easy and happy during my stay in Napoli.

I would like to thank to Prof. Bianca Maria Veneziani for her great company in the lab.

I'm grateful to all the teachers involved in "Computational Biology and Bioinformatics" PhD program. Their valuable lessons helped me to be introduced with a new realm of science.

I would like to thank my colleagues Roberto Amato and Giovanni Scala who helped me in many ways either in lab and whenever I need.

I would also like to thank Michele Pinelli, Micaela Montanari, Imma Castaldo, Fabio Acquaviva for their assistance and great company in the lab.

I'm grateful to all the past and present members of my lab, Sara corvigno, Agustina Nandone, Paola Vergara, Alessandra Cianflone, Luigi Copolla, Maria Letizia Cataldo for whom I got not only a nice environment in the lab, also their help and pleasant company made my life easy and joyful during my stay in Napoli, Italy. I shall always remember you all!

I would like to thank my husband, Sorif, for his love, patience, sincerity and support that have been great inspiration for me. Most importantly, I want to thank from core of my heart to my parents, brothers, father and mother-in-laws, sister-in-law and relatives for their endless love, well wishes and support that always fueled me in my way of life.

March 2013, Napoli