Stem Cell Reports

Article

OPEN ACCESS

Species-Specific Relationships between DNA and Chromatin Properties of CpG Islands in Embryonic Stem Cells and Differentiated Cells

Justin Langerman,[1,2,3] David Lopez,[2,3,4] Matteo Pellegrini,[2,3,4] and Stephen T. Smale[1,2,3,*]

[1]Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, CA 90095, USA
[2]Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA
[3]Broad Stem Cell Research Center, University of California, Los Angeles, CA 90095, USA
[4]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA 90095, USA
*Correspondence: smale@mednet.ucla.edu

https://doi.org/10.1016/j.stemcr.2021.02.016

SUMMARY

CpG islands often exhibit low DNA methylation, high histone H3 lysine 4 trimethylation, low nucleosome density, and high DNase I hypersensitivity, yet the rules by which CpG islands are sensed remain poorly understood. In this study, we first evaluated the relationships between the DNA and the chromatin properties of CpG islands in embryonic stem cells using modified bacterial artificial chromosomes. Then, using a bioinformatic approach, we identified strict CpG-island density and length thresholds in mouse embryonic stem and differentiated cells that consistently specify low DNA methylation levels. Surprisingly, the human genome exhibited a dramatically different relationship between DNA properties and DNA methylation levels of CpG islands. Further analysis allowed speculation that this difference is accommodated in part by evolutionary changes in the nucleotide composition of orthologous promoters. Thus, a change in the rules by which CpG-island properties are sensed may have co-evolved with compensatory genome adaptation events during mammalian evolution.

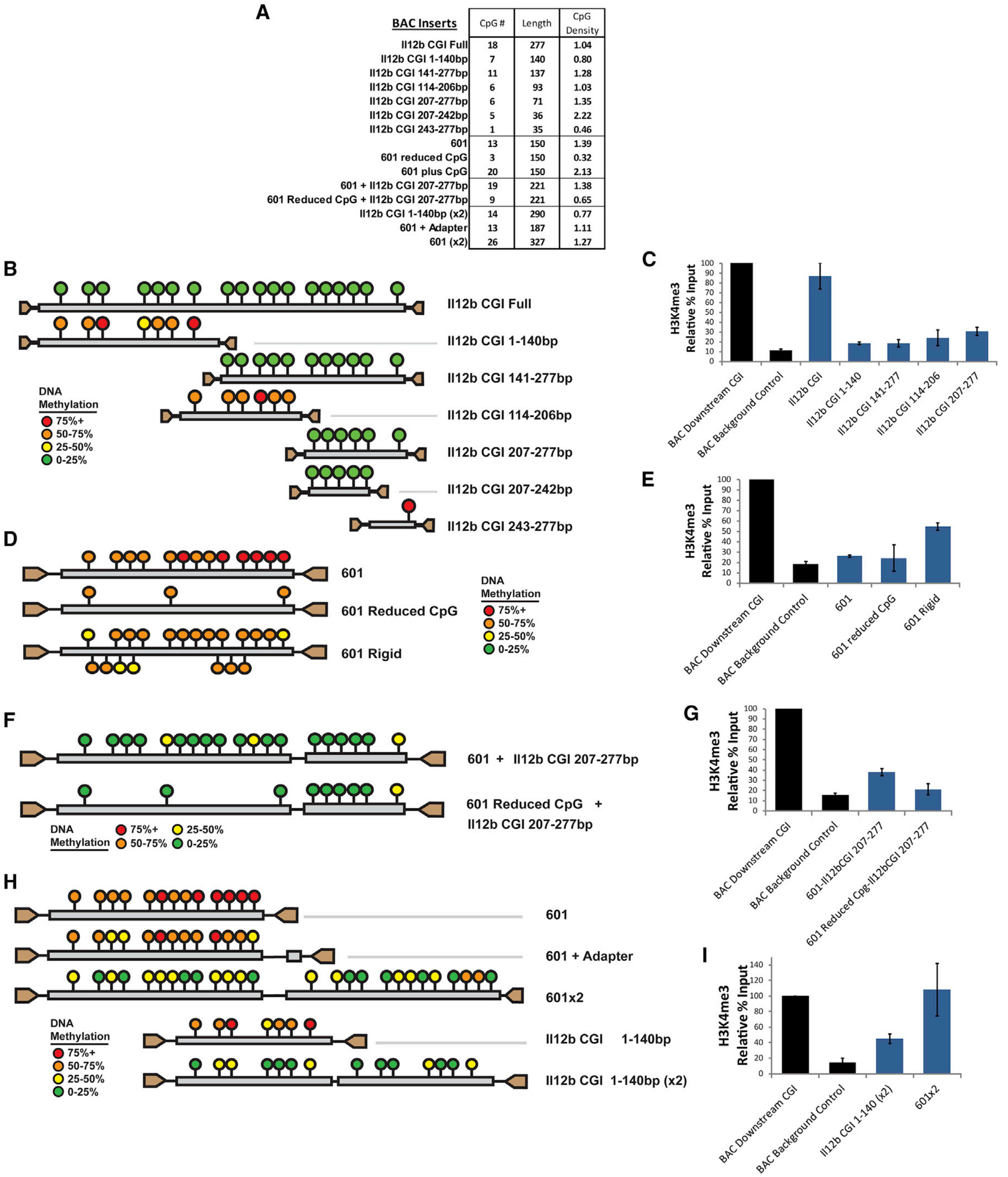
INTRODUCTION

Two fundamental goals of the eukaryotic gene regulation field are to understand how specific chromatin features are acquired at defined regions of the genome and to determine how these features help orchestrate the proper regulation of gene expression. Many chromatin features, including histone and DNA modifications, are regulated by enzymes recruited by transcription factors that bind DNA in a sequence-specific manner. However, chromatin structure can also be influenced by the nucleotide composition of a DNA region, or by molecular processes such as DNA replication, transcription, and DNA repair (Segal et al., 2006; Deaton and Bird, 2011; Morrison and Shen, 2009).

In mammals, CpG islands serve as an example of the impact of intrinsic nucleotide composition on chromatin structure (Deaton and Bird, 2011). CpG islands are found at approximately 70% of mammalian promoters and have been defined using different CpG dinucleotide prevalence criteria (Irizarry et al., 2009; Wu et al., 2010; Yu et al., 2017). Most definitions consider the length of the CpG-rich region (CGR), GC content, and the ratio of the observed density of CpG dinucleotides to the density expected if CpGs were distributed randomly (obs/exp ratio). Regardless of the criteria used, CpG islands often exhibit low DNA methylation, low nucleosome density, high DNase I hypersensitivity, and detectable histone H3 lysine 4 trimethylation (H3K4me3) (Bock et al., 2007). When associated with poised or silent genes, CpG islands can also exhibit high H3K27me3 (Bernstein et al., 2006).

Experimental evidence supports the hypothesis that the intrinsic DNA properties of CpG islands play an intimate role in specifying their characteristic chromatin properties (Lövkvist et al., 2016). The discovery of a protein motif that recognizes unmethylated CpG dinucleotides, termed the CXXC motif, uncovered one strategy by which the presence of several unmethylated CpGs in a CpG island might be sensed (Clouaire et al., 2012; Xu et al., 2018). CXXC domains can bind directly to and protect unmethylated CpG dinucleotides (Cierpicki et al., 2010). Studies have suggested that the sensing of intrinsic CpG island properties has functional impacts on transcription and development (Agarwal and Shendure, 2020; Hartl et al., 2019). However, the precise rules by which the DNA properties of CpG islands are sensed by CXXC motif-containing proteins or other chromatin regulators remain unknown. For example, is a defined density of CpG dinucleotides within a CGR of defined length essential for an impact on chromatin structure?

The intrinsic instability of nucleosomes assembled at CpG islands may also contribute to their characteristic chromatin properties. Early biochemical experiments demonstrated that GC base pairs help stabilize nucleosomes, but nucleosome stability is compromised in the absence of properly positioned AT dinucleotides, which provide the flexibility needed for DNA to wrap around the histone octamer (Drew and Travers, 1985; Lowary and Widom, 1998). *In vitro* biochemical studies and genome-wide nucleosome mapping studies have confirmed that CpG islands often exhibit intrinsic nucleosome instability (Fenouil et al., 2012; Ramirez-Carrozzi



BAC Inserts	CpG #	Length	CpG Density
Il12b CGI Full	18	277	1.04
Il12b CGI 1-140bp	7	140	0.80
Il12b CGI 141-277bp	11	137	1.28
Il12b CGI 114-206bp	6	93	1.03
Il12b CGI 207-277bp	6	71	1.35
Il12b CGI 207-242bp	5	36	2.22
Il12b CGI 243-277bp	1	35	0.46
601	13	150	1.39
601 reduced CpG	3	150	0.32
601 plus CpG	20	150	2.13
601 + Il12b CGI 207-277bp	19	221	1.38
601 Reduced CpG + Il12b CGI 207-277bp	9	221	0.65
Il12b CGI 1-140bp (x2)	14	290	0.77
601 + Adapter	13	187	1.11
601 (x2)	26	327	1.27

Figure 1. Experimental Analysis of CpG-Rich Sequences in a Gene-Desert BAC

(A) The table displays the basic DNA properties of CpG-rich sequences inserted into a gene-desert BAC, which was then pre-methylated and introduced into mouse ESCs.

(B) The diagrams represent the DNA methylation status in ESCs at inserts of the *Il12b* CpG island and deletion variants. Each circle represents a particular CpG position, the color of which represents the average ratio of methylated to unmethylated CpGs determined by bisulfite sequencing across multiple clones. The data shown for each insert are an aggregate of three or four clones. The key shows the color that corresponds to each DNA methylation level.

(C) The graph shows qPCR results from H3K4me3 ChIP DNA with primers specific to select inserts from (B). The signal is relative to the input fraction and normalized to a native unmethylated downstream CpG island on the transgenic BAC. The signal at a control region 2 kb

(legend continued on next page)

et al., 2009), raising the possibility that this instability contributes to the acquisition of the other chromatin properties that characterize CpG islands.

Although the DNA properties of CpG islands have the potential to contribute to their characteristic chromatin features, elucidation of this relationship has been confounded by the presence of transcription factor binding sites in CpG-island promoters. In fact, transcription factors have been shown to play a major role in dictating the chromatin properties of CpG islands through their ability to help maintain low CpG methylation levels and recruit enzymes involved in H3K4me3 deposition (Krebs et al., 2014). Is it possible that transcription factor binding is fully responsible for the chromatin properties of CpG islands, or do their DNA properties make an additional contribution?

To increase our understanding of CpG islands, we first evaluated DNA methylation at representative CpG islands and CpG-island variants in the context of bacterial artificial chromosomes (BACs) stably integrated into mouse embryonic stem cells (ESCs). We then carried out a systematic bioinformatic analysis of the relationships between the various DNA properties of CGRs and their characteristic chromatin properties. We also asked whether we could identify CpG-density and CGR-length thresholds that consistently coincided with low DNA methylation levels throughout the genome. This analysis led to a number of insights, the most striking of which was the existence of a dramatically different relationship between DNA properties and the DNA methylation state of the mouse and human genomes. Further analysis led to speculation that the two species adapted to these differences through alterations in the nucleotide composition of orthologous promoters and through the use of distinct mechanisms to mark silent genes.

RESULTS

Experimental Examination of Relationship between CpG island and DNA Methylation in Mouse ESCs

To extend our understanding of the relationship between CpG islands and chromatin properties, we first took advantage of a previously described strategy (Mendenhall et al., 2010) in which defined CGRs corresponding to known or artificial CpG islands were inserted into a 136-kb gene-desert BAC. The modified BACs were then premethylated (to help ensure broad methylation in the absence of activities promoting selective loss of methylation) and stably introduced into ESCs, followed by an analysis of DNA methylation in individual stably transfected clones by bisulfite sequencing; the goal was to examine the impact of DNA features on DNA methylation in a BAC system in which the DNA sequences could be readily manipulated. The properties of the 15 sequences inserted into the BAC for the current analysis are summarized in Figure 1A.

We first used this assay to analyze a small, 277-bp CGR of unknown function located 1 kb upstream of the mouse *Il12b* gene; this sequence is representative of a non-promoter CGR at the low end of the length and CpG density spectrum (see below) while exhibiting low DNA methylation and a significant H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq) peak in mouse ESCs (Vincent et al., 2013). When analyzed by bisulfite sequencing and H3K4me3 ChIP in the context of the gene-desert BAC stably integrated into the mouse ESC genome, this *Il12b* fragment was found to lack DNA methylation and to exhibit an H3K4me3 ChIP signal (Figures 1B and 1C). CpGs flanking the *Il12b* insert exhibited high DNA methylation and low H3K4me3 (data not shown).

We then analyzed six sub-fragments of this *Il12b* sequence to determine whether its DNA methylation and H3K4me3 properties are dictated by its overall DNA properties or by a specific DNA element. The results revealed that a 36-bp fragment (207–242) was necessary and sufficient to establish the unmethylated DNA state (Figure 1B), consistent with a model in which a transcription factor binding site rather than the DNA properties of the 277-bp fragment was responsible for the low DNA methylation. The identity of the factor responsible for the low DNA methylation remains unknown, but a near-consensus binding site for CTCF, a known mediator of low DNA methylation, is

upstream of the insertion site is also shown. The black bars represent the standard error from two immunoprecipitation replicates of at least two clones.

(D) The diagrams show the DNA methylation status at transgenic inserts of the 601 sequence and variants, similar to (B).

(E) The graph shows qPCR results from ChIP for H3K4me3 with primers specific to the 601, “601 reduced CpG,” and “601 plus CpG” variant inserts in ESCs. Normalization and controls are as in (C).

(F) The diagrams show DNA methylation levels determined by bisulfite sequencing at transgenic inserts of either 601 or “601 reduced CpG” fused to the *Il12b* CpG island 207–277 bp sequence from (B).

(G) The graph shows qPCR results from H3K4me3 ChIP DNA with primers specific to the inserts of 601 and “601 reduced CpG” fusion variant fusions to the *Il12b* CpG island 207–277 bp fragment.

(H) The diagrams show DNA methylation levels at transgenic inserts of tandem 601 (601 (×2)) or *Il12b* 1–140 bp fragments (*Il12b* 1–140 bp (×2)). Also shown are DNA methylation levels at an insert composed of 601 plus the adaptor used to adjoin tandem sequences.

(I) The graph shows qPCR results from H3K4me3 ChIP DNA with primers specific to the tandem insertions in (H).

present within this 36-bp sequence (Feldmann et al., 2013). Consistent with a possible role for CTCF, deletions of 9 or 15 bp within the CTCF binding site resulted in a substantial increase in DNA methylation in the context of the full-length CpG island (data not shown). Notably, only the intact, full-length 277-bp fragment acquired high H3K4me3 (Figure 1C), suggesting that additional transcription factors and/or specific DNA properties are needed for acquisition of this modification.

To further explore the possibility that DNA properties of CGRs might influence acquisition of the characteristic chromatin properties of CpG islands, we focused on a synthetic 150-bp CGR known as 601, which was originally identified as a sequence capable of assembling into a highly stable nucleosome (Lowary and Widom, 1998); this property is attributable to the fact that it combines a high GC content with properly spaced AT dinucleotides to allow optimal wrapping around the histone octamer. When analyzed in the context of the gene-desert BAC, 601 DNA remained heavily methylated and lacked H3K4me3 (Figures 1D and 1E), indicating that 13 CpG dinucleotides in a 150-bp fragment are insufficient for acquisition of low DNA methylation and high H3K4me3. We also tested two variants of 601; in one variant, “601 reduced CpG,” 10 of the 13 CpGs were converted to other dinucleotides, GpC or GpG, thereby retaining only 3 CpGs. In the other variant, “601 plus CpG,” 7 GpCs were converted to CpGs, thereby increasing the CpG density to an obs/exp ratio of 2.13, which exceeds the density of almost all CGRs in the human genome. Interestingly, the CpGs in both constructs remained heavily methylated when tested in the context of the gene-desert BAC, and H3K4me3 levels remained low (Figures 1D and 1E). Thus, altering CpG numbers and CpG density in a short DNA fragment appears insufficient to drive the loss of DNA methylation and acquisition of high H3K4me3.

Importantly, two different types of alterations to the 601 insert led to loss of DNA methylation. First, insertion of the 70-bp *Il12b* fragment that supports loss of DNA methylation adjacent to 601 promoted the efficient loss of methylation throughout the 601 sequence (Figure 1F), suggesting that the presence of an appropriate transcription factor binding site adjacent to a CpG-rich fragment is sufficient to promote loss of methylation. This spreading activity was not dependent on the high CpG density in 601, as similar loss of methylation was observed when the *Il12b* sequence was inserted adjacent to the “601 reduced CpG” variant (Figure 1F). The spreading activity also did not require placement of the *Il12b* fragment immediately adjacent to 601, as reduced 601 methylation was also observed when a 375-bp CpG-deficient BAC spacer fragment was inserted between 601 and the 70-bp *Il12b* fragment (data not shown). Interestingly, these constructs did not acquire appreciable H3K4me3 levels (Figure 1G), despite having CpG densities and lengths similar to those of the full-length *Il12b* CpG island.

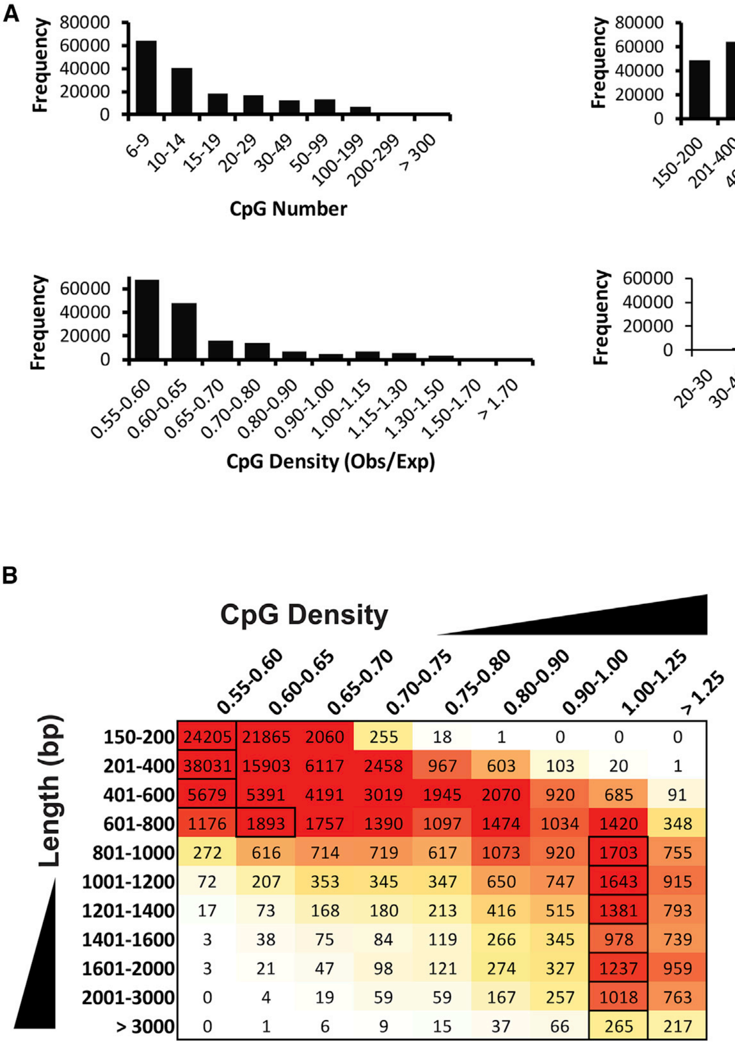
The second strategy that resulted in loss of DNA methylation at 601 was duplication of the sequence. Despite the inability of a single copy of 601 to support loss of DNA methylation, we observed efficient loss of methylation at the tandem sequence termed “601 (×2),” as well as acquisition of high H3K4me3 (Figures 1H and 1I). An insert containing only one copy of 601 adjoining the linker region used in the tandem version (“601 + adaptor”) retained high DNA methylation, demonstrating that the linker DNA does not contain a cryptic transcription factor binding site to promote methylation loss (Figure 1H). Although nucleosome density was difficult to accurately measure with these short sequences, these results also show that a CGR that is capable of assembling into stable nucleosomes can still readily acquire low DNA methylation and high H3K4me3.

A similar result was obtained when the *Il12b* CpG island sub-fragment, 1–140 bp, was duplicated; unlike the single insertion of this sequence, the duplicated version efficiently lost its methylation, with moderately increased H3K4me3 (Figures 1H and 1I). Together, the data in this figure support a hypothesis in which both transcription factor binding sites and the DNA properties of CpG islands can contribute to the loss of DNA methylation and acquisition of H3K4me3.

Relationships between the DNA Properties of CpG-Rich Regions

To define further the relationship between the DNA properties of CpG islands and their characteristic chromatin properties, we developed a bioinformatic approach. We first defined a collection of CGRs within the human genome using low-stringency criteria. Most previous studies used higher stringency criteria to capture primarily those DNA segments that function as CpG islands (Irizarry et al., 2009; Wu et al., 2010; Yu et al., 2017). In contrast, a large fraction of the DNA segments within our collection lack CpG-island properties and functions, thereby facilitating a detailed examination of the relationships between DNA properties and chromatin properties.

Specifically, we selected all non-repetitive DNA segments of at least 150 bp that possess an obs/exp ratio for CpG dinucleotides (i.e., CpG density) of at least 0.55. If two or more overlapping 150-bp segments met this density criterion, they were merged into one longer segment. We did not normalize for GC percentage when calculating CpG density; normalization would be based on an uncertain assumption that GC content does not contribute to regulatory functions. Using these criteria, 173,307 CGRs were identified, comprising approximately 5% of the non-



Length (bp)	0.55-0.60	0.60-0.65	0.65-0.70	0.70-0.75	0.75-0.80	0.80-0.90	0.90-1.00	1.00-1.25	> 1.25
150-200	24205	21865	2060	255	18	1	0	0	0
201-400	38031	15903	6117	2458	967	603	103	20	1
401-600	5679	5391	4191	3019	1945	2070	920	685	91
601-800	1176	1893	1757	1390	1097	1474	1034	1420	348
801-1000	272	616	714	719	617	1073	920	1703	755
1001-1200	72	207	353	345	347	650	747	1643	915
1201-1400	17	73	168	180	213	416	515	1381	793
1401-1600	3	38	75	84	119	266	345	978	739
1601-2000	3	21	47	98	121	274	327	1237	959
2001-3000	0	4	19	59	59	167	257	1018	763
> 3000	0	1	6	9	15	37	66	265	217

Figure 2. Nucleotide Properties of Human CGRs

(A) The bar graph shows the distribution of the frequency of all 173,307 human CGRs within bins for total CpG number, total length, average CpG density, or average GC content.

(B) A frequency histogram shows the CGR distribution among total length bins and CpG density bins. Red denotes high occurrence. For each length bin, the highest co-occurring CpG density bin is indicated (black box).

repetitive portion of the human genome. As expected, broad ranges of CpG numbers, densities, CGR lengths, and GC percentages were observed within the pool of CGRs (Figure 2A).

From an analysis of the relationships between the various DNA properties of the CGRs, the most striking finding was a dramatic difference in the CpG density distribution as a function of CGR length (Figure 2B). CGRs between 150 and 600 bp in length frequently possess CpG densities at the low end of the range used in the analysis (0.55–0.60). However, CGRs greater than 800 bp in length were found to be distributed around a much higher CpG density peak of 1.00–1.25, with surprisingly few CGRs near the density minimum (Figure 2B). The infrequent occurrence of long CGRs with low average CpG densities is consistent with a hypothesis in which long CGRs have remained unmethylated in germ cell lineages with such high levels of stability that they have had little opportunity for loss of CpG dinucleotides during evolution via 5-methylcytosine deamination (Cohen et al., 2011).

Relationships between Chromatin Properties of CGRs

As an initial step toward evaluating chromatin properties, published datasets from human ESCs (see Experimental Procedures) were analyzed, revealing the expected distribution of each chromatin property among the set of CGRs

A

			Mouse ESC (v6.5)			Mouse ESC (E14)			Mouse Frontal Cortex			Mouse Mammary			Mouse Kidney			Mouse Heart		
	CpG Density (obs/ex)	CGR Length (bp)	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth
Promoters	>1.0	>1,000	0	5154	0.0	0	5150	0.0	5	5152	0.1	5	5180	0.1	5	5160	0.1	5	5163	0.1
	0.8-1.0	>1,000	3	1476	0.2	0	1476	0.0	9	1474	0.6	7	1488	0.5	9	1481	0.6	11	1483	0.7
	>1.0	600-1,000	0	2591	0.0	1	2578	0.0	8	2594	0.3	10	2601	0.4	10	2591	0.4	8	2593	0.3
	0.8-1.0	600-1,000	4	1514	0.3	6	1495	0.4	30	1515	2.0	25	1538	1.6	28	1513	1.9	24	1509	1.6
Exons	>1.0	>1,000	1	322	0.3	0	319	0.0	30	322	9.3	36	319	11.3	34	319	10.7	29	319	9.1
	0.8-1.0	>1,000	21	245	8.6	15	243	6.2	97	245	39.6	94	244	38.5	106	244	43.4	95	244	38.9
	>1.0	600-1,000	0	253	0.0	1	251	0.4	29	253	11.5	37	253	14.6	35	253	13.8	28	253	11.1
	0.8-1.0	600-1,000	33	439	7.5	25	438	5.7	174	439	39.6	172	439	39.2	186	439	42.4	181	439	41.2
Introns	>1.0	>1,000	0	93	0.0	0	88	0.0	0	93	0.0	1	91	1.1	0	90	0.0	0	90	0.0
	0.8-1.0	>1,000	1	69	1.4	1	67	1.5	4	70	5.7	2	72	2.8	4	68	5.9	3	68	4.4
	>1.0	600-1,000	0	100	0.0	1	94	1.1	4	100	4.0	2	102	2.0	7	100	7.0	5	100	5.0
	0.8-1.0	600-1,000	14	151	9.3	12	146	8.2	27	151	17.9	20	153	13.1	28	150	18.7	28	151	18.5
Intergenic	>1.0	>1,000	0	299	0.0	0	275	0.0	1	303	0.3	2	309	0.6	4	291	1.4	3	295	1.0
	0.8-1.0	>1,000	4	227	1.8	5	216	2.3	13	224	5.8	10	229	4.4	13	216	6.0	12	217	5.5
	>1.0	600-1,000	2	324	0.6	3	317	0.9	7	325	2.2	4	336	1.2	9	323	2.8	4	325	1.2
	0.8-1.0	600-1,000	19	440	4.3	14	426	3.3	50	443	11.3	43	460	9.3	52	440	11.8	53	440	12.0

B

			Human ESC (HUES64)			Human ESC (H1)			Human Frontal Cortex			Human Hair Follicle			Human PBL			Human FB (IMR90)		
	CpG Density (obs/ex)	CGR Length (bp)	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	No. Meth	Total	% Meth
Promoters	>1.0	>1,000	149	8079	1.8	105	8076	1.3	76	8141	0.9	66	8147	0.8	93	8147	1.1	67	8112	0.8
	0.8-1.0	>1,000	168	1693	9.9	147	1692	8.7	124	1710	7.3	115	1720	6.7	149	1720	8.7	117	1709	6.8
	>1.0	600-1,000	121	2461	4.9	90	2454	3.7	45	2479	1.8	48	2480	1.9	65	2480	2.6	40	2479	1.6
	0.8-1.0	600-1,000	204	1404	14.5	165	1399	11.8	145	1428	10.2	137	1439	9.5	170	1439	11.8	128	1425	9.0
Exons	>1.0	>1,000	497	1077	46.1	437	1075	40.7	340	1070	31.8	368	1080	34.1	425	1080	39.4	355	1070	33.2
	0.8-1.0	>1,000	740	963	76.8	690	964	71.6	656	966	67.9	622	969	64.2	726	969	74.9	679	966	70.3
	>1.0	600-1,000	291	567	51.3	245	568	43.1	219	567	38.6	220	568	38.7	258	568	45.4	214	567	37.7
	0.8-1.0	600-1,000	920	1150	80.0	859	1150	74.7	824	1160	71.0	782	1165	67.1	914	1164	78.5	842	1160	72.6
Introns	>1.0	>1,000	140	407	34.4	131	408	32.1	115	412	27.9	121	415	29.2	133	415	32.0	120	412	29.1
	0.8-1.0	>1,000	251	406	61.8	240	406	59.1	207	404	51.2	206	410	50.2	244	409	59.7	209	404	51.7
	>1.0	600-1,000	141	339	41.6	126	339	37.2	100	338	29.6	101	341	29.6	131	341	38.4	103	338	30.5
	0.8-1.0	600-1,000	334	577	57.9	322	577	55.8	272	584	46.6	267	585	45.6	324	584	55.5	273	584	46.7
Intergenic	>1.0	>1,000	318	1339	23.7	265	1338	19.8	174	1393	12.5	195	1402	13.9	225	1402	16.0	179	1392	12.9
	0.8-1.0	>1,000	370	844	43.8	315	835	37.7	267	862	31.0	259	864	30.0	297	864	34.4	238	861	27.6
	>1.0	600-1,000	242	900	26.9	186	904	20.6	121	925	13.1	130	925	14.1	161	925	17.4	125	925	13.5
	0.8-1.0	600-1,000	576	1213	47.5	510	1215	42.0	422	1249	33.8	404	1257	32.1	491	1258	39.0	389	1248	31.2

C

			Mouse ES Average			Mouse Tissue Average			Human ES Average			Human Tissue Average		
	CpG Density (obs/ex)	CGR Length (bp)	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth	# Meth	Total	% Meth
Promoters	>1.0	>1,000	0	5152	0.0	5	5164	0.1	127	8078	1.6	76	8137	0.9
	0.8-1.0	>1,000	2	1476	0.1	9	1482	0.6	158	1693	9.3	126	1715	7.4
	>1.0	600-1,000	1	2585	0.0	9	2595	0.3	106	2458	4.3	50	2480	2.0
	0.8-1.0	600-1,000	5	1505	0.3	27	1519	1.8	185	1402	13.2	145	1433	10.1
Exons	>1.0	>1,000	1	321	0.2	32	320	10.1	467	1076	43.4	372	1075	34.6
	0.8-1.0	>1,000	18	244	7.4	98	244	40.1	715	964	74.2	671	968	69.3
	>1.0	600-1,000	1	252	0.2	32	253	12.7	268	568	47.2	228	568	40.1
	0.8-1.0	600-1,000	29	439	6.6	178	439	40.6	890	1150	77.3	841	1162	72.3
Introns	>1.0	>1,000	0	91	0.0	0	91	0.3	136	408	33.3	122	414	29.6
	0.8-1.0	>1,000	1	68	1.5	3	70	4.7	246	406	60.5	217	407	53.2
	>1.0	600-1,000	1	97	0.5	5	101	4.5	134	339	39.4	109	340	32.0
	0.8-1.0	600-1,000	13	149	8.7	26	151	17.0	328	577	56.8	284	584	48.6
Intergenic	>1.0	>1,000	0	287	0.0	3	300	0.8	292	1339	21.8	193	1397	13.8
	0.8-1.0	>1,000	5	222	2.0	12	222	5.4	343	840	40.8	265	863	30.7
	>1.0	600-1,000	3	321	0.8	6	327	1.8	214	902	23.7	134	925	14.5
	0.8-1.0	600-1,000	17	433	3.8	50	446	11.1	543	1214	44.7	427	1253	34.0

Figure 3. Prevalence of Highly Methylated CGRs Identified Using Defined CpG Density/CGR Length Criteria in Mouse and Human ESC and Differentiated Cells

(A and B) The tables show the counts of CGRs in mouse and human DNA methylation datasets that meet the DNA property criteria labeled at left. CGRs are separated by genomic location and by non-overlapping CpG density and length ranges. Cell line names are in parentheses, with all other datasets derived from primary tissues. For all six mouse (A) and six human (B) datasets, the number of CGRs with high DNA methylation (>70%) in each criteria range is shown under "# Meth" next to the number of regions that qualify for the criteria, "Total." The last column for each methylome group is the percentage of CGRs methylated for each criterion, "% Meth." The tables are colored by frequency; increased gray indicates higher CGR numbers, while increased red indicates a higher percentage in the "% Meth" column. CGRs

(legend continued on next page)

(Figure S1). Of greatest relevance for the current analysis, a bimodal distribution of DNA methylation was observed in the ESCs, with 17% and 77% of CGRs exhibiting either low (0%–30%) or high (70%–100%) DNA methylation levels, and only 6% exhibiting intermediate levels.

The datasets were then used to examine relationships between the chromatin properties. As expected, these results revealed close correlations between low CpG methylation, high or medium H3K4me3, and high or medium DNase HS (Figure S2). In contrast, nucleosome density and H3K27me3 did not correlate with these chromatin properties (Figure S2). Overall, the results support a model in which the acquisition of histone H3K4me3, DNase HS, and low DNA methylation at CGRs relies on closely linked mechanisms. In contrast, distinct mechanisms may be responsible for the broad acquisition of low nucleosome density and measurable histone H3K27me3.

The modest correlation between DNase HS and nucleosome density (derived from micrococcal nuclease sequencing [MNase-seq] data) was initially surprising, given that both assays provide a measure of the physical accessibility of chromatin to nuclease cleavage. However, this lack of correlation is consistent with a model in which the DNase HS assay monitors the eviction of one or more nucleosomes from a specific portion of a CGR (leading to hypersensitivity to one or more focused DNase I cleavage events), whereas the MNase assay provides an average nucleosome density throughout the entire CGR. In fact, subsequent analyses showed that nucleosome density correlates most closely with GC percentage rather than the chromatin properties associated with active transcription or with transcription itself (data not shown). Thus, we speculate that, although GC percentage may be a major driver of overall nucleosome density within a CGR, this property may have little relevance to transcriptional control.

Further analysis of genomic location revealed that CGRs at promoters exhibited much lower DNA methylation levels, higher H3K4me3, and higher DNase HS than non-promoter CGRs (Figure S3). Promoter CGRs also exhibited higher CpG densities and were generally longer than non-promoter CGRs (Figure S3). Thus, the three properties associated with active chromatin correlate with DNA properties, but also with regions (i.e., promoters) that are known to be transcriptionally active. Subsequent analyses, in which promoter and non-promoter CGRs were separated into bins based on their DNA properties, with the chromatin properties of each bin then examined, yielded results consistent with the notion that both DNA properties and transcription factor binding influence the chromatin properties of CGRs (data not shown). Importantly, repetition of the above analysis with mouse ESCs yielded similar results (data not shown).

DNA Properties in Mouse ESCs that Consistently Coincide with Low DNA Methylation

The data presented above are consistent with models in which the chromatin properties of CpG islands are influenced by both intrinsic DNA properties and transcription factor binding. However, these results consist largely of statistical trends rather than precise rules. We therefore asked whether specific DNA properties could be identified that consistently predict a defined chromatin state, with the analysis first performed with mouse CGRs. We focused on DNA methylation because of the bimodal distribution of this chromatin property, which allowed reliable quantitation and greater consistency among datasets.

When considering CpG density and CGR length simultaneously in a mouse ESC dataset, we first noted that high DNA methylation was not observed at any of the 5,154 promoter CGRs, and at only 1 of the 714 non-promoter CGRs, exhibiting a CpG density >1.0 and a CGR length >1 kb (Figure 3A, mouse ESC v.6.5 data, >1 CpG density/>1,000 bp length rows). Notably, the 5,154 promoter CGRs that fulfill these criteria represent approximately 21% of all RefSeq promoters. When the CGR length stringency was arbitrarily reduced to 600 bp, 2,591 additional promoter and 677 non-promoter CGRs were added to the pool, yet none of the additional promoters and only 2 of the non-promoter CGRs exhibited high DNA methylation (Figure 3A, mouse ESC v.6.5 data, >1 CpG density/600–1,000 bp rows). An arbitrary reduction in the CpG density stringency to 0.8, while maintaining the CGR length stringency of 1 kb, or simultaneous reductions in the length and density stringencies to 600 bp and 0.8, respectively, resulted in only gradual increases in the percentage of methylated CGRs (Figure 3A, mouse ESC v.6.5 data). Examination of additional bins showed that the prevalence of high DNA methylation increases as the CpG density and CGR length stringencies are reduced (data not shown). Importantly, highly similar profiles were observed with a second independent mouse ESC DNA methylation dataset (Figure 3A, mouse ESC E14 data [please note that the total number of CGRs stated differs slightly among datasets due to differences in CGRs exhibiting sequencing reads]).

Thus, a strict rule appears to exist in mouse ESCs by which, with only one exception out of 5,868 CGRs, a

with insufficient bisulfite sequencing reads were discarded, which accounts for the slight variation in qualifying CGRs between methylomes. In addition, promoter CGR classifications were confirmed manually for all methylated mouse CGRs shown and all methylated human CGRs >1.0 CpG density and >1 kb CGR length (see Figure 4). PBL, peripheral blood leukocytes; FB, fibroblasts.
(C) Averaged and rounded values for the mouse and human ESC and somatic cell data from (A) and (B) are shown.

CpG density >1.0 combined with a CGR length >1 kb coincides with an absence of high methylation. This strict rule breaks down gradually as density and length stringencies are reduced. Notably, the small number of CGRs that were methylated in both of the mouse ESC lines are biased toward exons and introns (Figure 3A). This finding is consistent with evidence that Dnmt3b is recruited to gene bodies during active transcription to promote DNA methylation (Baubec et al., 2015; Morselli et al., 2015).

A Competition Model May Explain the DNA Methylation State of CGRs

An examination of the mouse ESC profiles in Figure 3A seems most consistent with a model in which the DNA methylation state of a CGR is dictated by competition between mechanisms that promote an unmethylated state and opposing mechanisms that promote a methylated state. In this speculative model, CpG density and CGR length help promote an unmethylated state, perhaps through their ability to bind Cfp1 or other unmethylated CpG binding proteins (Cierpicki et al., 2010; Clouaire et al., 2012). At CGRs that exceed a CpG density of 1.0 and a length of 1 kb, such a mechanism may be sufficient to ensure an unmethylated state. Transcription factors may also help promote the unmethylated state at CGRs that exceed these density and length thresholds.

In opposition to these two mechanisms, transcription-directed Dnmt3b recruitment is likely to promote DNA methylation at a subset of transcriptionally active gene bodies. Methylation is also likely to be promoted by other mechanisms, such as spreading from flanking regions of low CpG density. According to this model, as the CpG density and CGR length stringencies are reduced, a larger percentage of CGRs are found to be methylated, with the highest percentage of methylated CGRs in gene bodies (due to transcription-coupled Dnmt3b recruitment), the lowest percentage of methylated CGRs at promoters (due to transcription factor-mediated protection from methylation), and an intermediate percentage of methylated CGRs at intergenic regions.

An Altered Competitive Balance in Differentiated Cells

To extend our analysis, we analyzed DNA methylation datasets from four independent differentiated cell types from mice. Interestingly, in all of these datasets, a larger number of CGRs with high methylation were observed in all genomic locations and in all CGR density/length bins in comparison with the two independent mouse ESC datasets (Figure 3A, promoter data summarized in Figure 4A). For example, although no promoters with a CpG density >1.0 and CGR length >1 kb were methylated in ESCs, five promoters in this stringency bin were methylated in all four differentiated cell types (Figure 4A).

Interestingly, the same five promoters were methylated in all four cell types and all were found to be associated with sperm-specific genes (Figure 4A), providing initial evidence that somatic cells contain a mechanism to promote the methylation of germ-lineage-specific promoters possessing CpG criteria that typically would promote an unmethylated state. When the CpG density and CGR length stringencies were reduced, additional methylated promoter CGRs were detected in the four datasets from mouse differentiated cells, with a strong bias toward promoters of germ-lineage genes (Figures 3A and 4A). The acquisition of DNA methylation at the promoters of germ-lineage genes appears to act during the maturation of ESCs to epiblasts, as these same promoters were found to be highly methylated in a mouse epiblast reduced-representation bisulfite sequencing dataset (data not shown). These findings are consistent with a study that documented increased methylation at the CpG-island promoters of germ-lineage genes following ESC maturation (Auclair et al., 2014). More generally, the increased DNA methylation observed in somatic cells in comparison with ESCs is consistent with abundant evidence that chromatin structure is in a more open state in ESCs.

Shifting our attention to non-promoter CGRs in the four differentiated cell types, between 31 and 39 non-promoter CGRs exceeding a density of 1.0 and length of 1 kb were found to be highly methylated in comparison with only one in the ESC datasets, with the vast majority of these methylated non-promoter CGRs found in exons (Figure 3A). The prevalence of methylated CGRs in exons in comparison with introns appears to be due to a bias toward methylation of CGRs near the 3′ end of genes, which tends to coincide with exons (i.e., 3′ untranslated regions) rather than introns (data not shown). Similar to the observations in ESCs, the percentage of methylated CGRs increased gradually with reduced CpG stringencies (Figure 3A).

Most notably, increases in the percentages of methylated CGRs increased concordantly at promoters, exons, introns, and intergenic regions in the reduced stringency bins, when the ESC profiles were compared with the somatic cell profiles. This finding is consistent with the competition model, as the balance appears to have shifted in these differentiated cells toward mechanisms that promote methylation and/or away from mechanisms that protect against methylation. Because the altered balance was observed at all genomic locations, we speculate that it is due to reduced potency of a putative mechanism that "measures" CpG density and CGR length to support an unmethylated state. If the altered balance were primarily due to enhanced recruitment of Dnmt3b to gene bodies or to

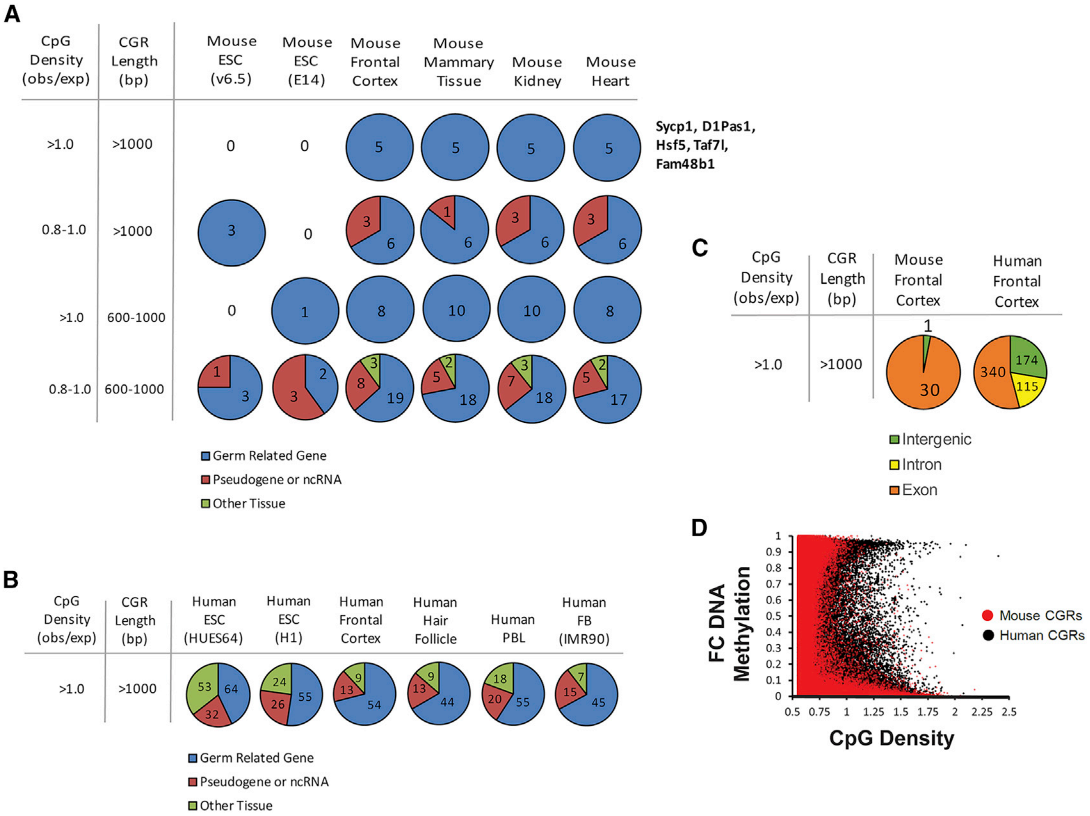


Figure 4. The Relationships between DNA Properties and DNA Methylation Differ in Mice and Humans and during Development

(A) The pie charts show the gene types associated with the highly methylated mouse promoter CGRs identified using the DNA property criteria at the left (see also Figure 3). Each graph shows the number of CGRs for each mouse cell type that have an associated gene that is either germline related, pseudogene/non-coding, or non-germ related (skin, neuronal, etc.). The five CGRs with CpG density >1.0 and length >1 kb and that have high methylation are from the same five genes in all four differentiated cell types; the names of these five genes are displayed to the right.

(B) The pie charts show the characterization of the confirmed promoter human CGRs with CpG density >1.0 and length >1 kb that are methylated from each cell type, as in (A).

(C) The pie charts show a comparison of the methylated mouse and human CGRs with CpG density >1.0 and length >1 kb, at non-promoter locations in frontal cortex. Each graph shows the number of methylated CGRs at exons, introns, or intergenic regions, colored by the key at the bottom.

(D) The dot plot shows the distribution of average DNA methylation in frontal cortex at all CGRs compared with CpG density, for both mouse (red) and human (black).

changes in transcription factor binding to promoters, one would expect changes to be observed primarily at gene bodies or promoters.

A Greatly Altered Competitive Balance in Human Cells

Most surprisingly, in human cells in comparison with mouse cells, the balance appears to be tilted much further in favor of the mechanisms promoting DNA methylation. This finding is apparent in two datasets from human ESCs and four datasets from human differentiated cell types (Figures 3B and 4B). Notably, the prevalence of methylated CGRs in human ESCs was comparable to that in human differentiated cell types (Figures 3B and 4B), consistent with the notion that human ESCs are thought to be more analogous to mouse post-epiblast cells than to mouse ESCs (see below). However, much higher percentages of CGRs in all stringency bins and at all genomic locations exhibited high methylation in the six human cell datasets compared with the four datasets derived from differentiated mouse cells (compare Figure 3A to Figure 3B and note averages in Figure 4C).

At human promoters within the highest stringency bin (CpG density >1.0 and CGR length >1.0), methylation was biased toward germ-lineage-specific genes, as in mice (Figure 4B). However, the number of methylated promoters in humans was much larger than in mice and was not

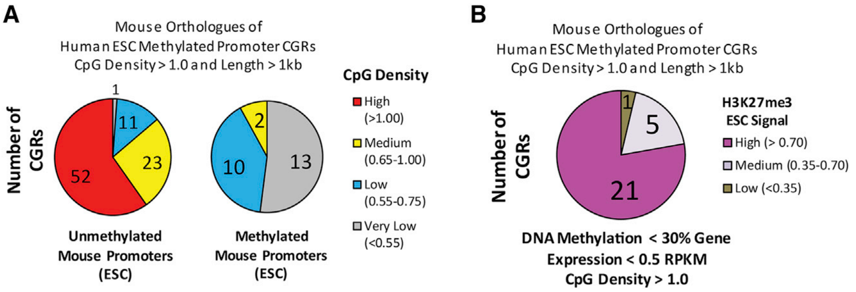


Figure 5. Species-Specific Adaptation of Promoters to Accommodate the Different Relationships between DNA Properties and DNA Methylation

(A) For this chart, 112 mouse promoter CGRs were selected that were orthologous to methylated human CGRs from ESCs with length >1 kb and CpG density >1.0. The graph to the left shows the corresponding 87 orthologous mouse promoters that were unmethylated in mouse ESCs, numbered and colored by CpG density as described in the key at right; 52 of these mouse ortholog promoters retain high CpG densities that are resistant to DNA methylation in mice. The right graph shows the 25 mouse promoters that were methylated in mouse ESCs similarly labeled and colored; the CpG densities of these promoters are consistently lower than those of their human orthologs, providing an explanation for their ability to be methylated in mouse cells. The very low (<0.55) CpG density group reflects genes for which a mouse ortholog exists, but which have no DNA region that fulfills the minimum criteria of a CGR within 500 bp of the transcription start site (TSS). DNA methylation at very low regions was calculated by averaging all CpGs within 500 bp of the TSS.

(B) Of the 52 mouse promoters from (A) that retain the high CpG densities observed with their human orthologs (and therefore are unmethylated in mice because CGRs with these DNA properties are almost never methylated), 27 were found to be silent in mouse ESCs (<0.5 RPKM). H3K27me3 levels are shown for these 27 promoters, suggesting that H3K27me3 may contribute to the silencing of these “methylation-resistant” mouse promoters.

limited to germ-lineage genes (compare Figures 4A and 4B). Also similar to the observations in mice, high methylation at human non-promoter locations was most prevalent within exons (Figures 3 and 4C). However, once again, the number of methylated exons was consistently much higher in humans than in mice. For example, mouse and human frontal cortex contained 31 and 629 methylated CGRs, respectively, in the highest stringency bin (Figure 4C). The scatterplot in Figure 4D, which compares DNA methylation and CpG density in mouse and human frontal cortex, further emphasizes the fact that high DNA methylation is observed at CGRs with high CpG densities in humans but not in mice.

As mentioned above, typical human ESC lines, unlike mouse ESC lines, possess molecular characteristics that are most similar to the post-epiblast stage of development, including higher global DNA methylation (Nichols and Smith, 2009). Recently it has been shown that by using certain culture conditions, ESCs can be converted from this primed phenotype to a naive, blastocyst-like stage (Guo et al., 2017; Takashima et al., 2014; Theunissen et al., 2014). By analyzing a DNA methylome generated by Guo et al. (2017) from these naive human ESCs, with a direct comparison with a methylome from the more mature primed cells, we observed that DNA methylation in the high-stringency CGR bins in the naive cells was lower than in the human primed cells (Figure S4), but remained substantially higher than the methylation levels in mouse ESCs (and more comparable to mouse differentiated tissues). Therefore, even at the most blastocyst-like stage with reduced overall DNA methylation, human ESCs maintain a skewed balance toward DNA methylation at CGRs in comparison with mouse ESCs.

Species-Specific Adaptation of Promoter Sequences

The striking difference between humans and mice with respect to the relationship between DNA properties and DNA methylation was unexpected and difficult to rationalize. In particular, one would expect such fundamental relationships to be well conserved through mammalian evolution. Instead, on the basis of the differences described above, many orthologous promoters are predicted to be highly methylated when silent in human cells, but not when silent in mouse cells.

To examine how this fundamental difference might be tolerated, we analyzed the properties of orthologous promoters. Of the human genes whose promoters exhibited a CpG density >1.0, a CGR length >1 kb, and high methylation in human ESCs, 112 had mouse orthologs that could be clearly identified. Strikingly, 25 of these 112 mouse promoters exhibited high methylation in mouse ESCs. On the basis of the data shown in Figure 3, high methylation would not be expected in mice if the mouse promoters possessed the same DNA properties as their human orthologs. Interestingly, each of these mouse promoters exhibited a CpG density that was much lower than that observed in the human ortholog (Figure 5A, right). Furthermore, the two methylated mouse promoters with the highest CpG densities (between 0.75 and 1.00) were shorter than their human orthologs (data not shown). Thus, all of these mouse promoters exhibited properties that are compatible with DNA methylation in mice. In other words, the CpG densities and CGR lengths of the mouse and human promoters appear to have adapted to their distinct, species-specific relationships between DNA properties and DNA methylation; the mouse orthologs possess reduced CpG densities and CGR lengths, allowing them to acquire the same high levels of DNA methylation when silent as their human counterparts. (To determine whether these results can be explained by frequent differences in CpG densities of orthologous promoters, we examined the human orthologs of mouse promoter CGRs in the highest CpG-density bin; 82% of the human orthologs exhibited a high CpG density [>1.0], demonstrating that differences in CpG density are relatively infrequent [data not shown].)

The mouse orthologs of the remaining 87 methylated human promoters were found to be unmethylated in mouse ESCs (Figure 5A, left). Thirty-five of these 87 mouse promoters exhibited reduced CpG densities, which may allow them to be silenced by DNA methylation in other cell types. However, the other 52 promoters exhibited high CpG densities in mice, just as in humans (Figure 5A, left). Most of these 52 promoter CGRs were also longer than 1 kb, suggesting that they may never be susceptible to DNA methylation in mice, even when silent, in contrast to the high methylation observed at their human orthologs. Twenty-seven of these 52 promoters were expressed below 0.5 RPKM in mouse ESCs. An examination of histone modifications at these 27 promoters revealed that all but one exhibited high or medium levels of H3K27me3 (Figure 5B). These results suggest that, despite CpG properties that render these mouse promoters resistant to DNA methylation (in contrast to their human orthologs), they remain susceptible to silencing via an H3K27me3-dependent mechanism. Thus, we speculate that different rules governing the relationship between DNA properties and DNA methylation in mice and humans can be tolerated because some orthologous promoters have altered their nucleotide compositions in a species-specific manner to maintain susceptibility to DNA methylation in both species; other promoters may rely exclusively on H3K27me3 to maintain a silent state in mice, with DNA methylation contributing to silencing of the human ortholog.

DISCUSSION

Through a systematic analysis of the relationships between the DNA and the chromatin properties of CGRs in the human and mouse genomes, we provide evidence in support of models in which both transcription factor binding and intrinsic DNA properties contribute to the regulation of the chromatin properties at CpG islands. Of greatest importance, an effort to uncover rules by which DNA properties influence DNA methylation revealed striking differences between mouse ESCs and mouse somatic cells, and between mice and humans, with respect to the intrinsic relationship between DNA properties and DNA methylation. We speculate that the genomes of the two species accommodated these differences through adaptation of their promoter sequences and of the modes of silencing employed at orthologous genes.

The profiles described throughout this analysis support a model in which mechanisms promoting the loss of DNA methylation at CpG islands (transcription factor binding and the recognition of CpG-rich DNA regions) compete with mechanisms that promote DNA methylation (transcription-coupled Dnmt3b recruitment and perhaps spreading from CpG-poor regions). The balance between these opposing mechanisms appears to shift during development and to differ in humans and mice. We propose that the shift in this balance is due to a change in the potency of the unknown mechanism by which CpG density and CGR length properties are sensed, since a change in this one property could explain the fact that the competitive balance is shifted at all genomic regions.

It is important to note that, although a surprisingly firm rule can be defined in mouse ESCs, in that only 1 of 5,868 CGRs with a CpG density >1.0 and CGR length >1 kb exhibits high methylation, the breakdown of this rule in reduced stringency bins is gradual. This observation suggests that more refined rules remain to be elucidated to explain why, for example, 30 exonic CGRs in frontal cortex (CpG density >1.0, length >1 kb) are methylated, whereas the remaining 292 exonic CGRs with similar density and length properties remain unmethylated. Interestingly, 3 of these 30 methylated exonic CGRs contain highly repetitive sequences (data not shown), which may promote the acquisition of DNA methylation. The remaining 27 also exhibit a tendency toward repetitive sequence content that differs substantially from that observed at the CGRs that remain unmethylated (data not shown). Additional studies are also needed to gain an understanding of the mechanism by which CpG density and CGR length are measured. As suggested above, variations in this mechanism may be responsible for the species-specific relationships between DNA properties and DNA methylation, thereby increasing the importance of future studies in this direction.

EXPERIMENTAL PROCEDURES

Identification and Analysis of CGRs

A script (see below) was written to scan the repeat-masked hg19 human genome and the mm9 mouse genome for CpG dinucleotide occurrence. Sliding 150-bp windows were scanned and selected if the obs/exp CpG ratio was greater than 0.55:

Obs/exp ratio = CpG number ÷ (size of window in bp × probability of random CpG [1/16]). CGRs were defined by specific genomic coordinates and overlapping regions were combined.

Distribution analyses were performed using frequency histograms bins in Microsoft Excel. Genomic locations of non-promoter regions (including genic regions) were defined by the UCSC Refseq database. Promoter locations were determined as overlap ± 500 bp of any transcription start site defined in the UCSC RefSeq database. Overlapping locations were called by the following hierarchy: promoter, UTR (omitted in Figure 3), exon, intron, intergenic.

Analysis of CGR Chromatin Properties

ChIP-seq datasets for H3K4me3 and H3K27me3 for H1 human ESCs and E14/Bruce4 mouse ESCs, and H3K4me3 data from post-mortem human frontal cortex, were obtained from the ENCODE database on the UCSC Genome Browser (Bernstein et al., 2010; Ernst et al., 2011; The Mouse ENCODE Consortium et al., 2014). The histone modification signal at CGRs was calculated by averaging of the ChIP sequencing signal across the CGR position interval. DNase HS data for H1, E14 human ESCs, and human frontal cortex were from publicly available datasets (Thurman et al., 2012; Vierstra et al., 2014). DNase HS signals at CGRs were calculated as the maximal peak score overlapped by the CGR interval. MNase nucleosome mapping data for human and mouse ESCs were downloaded from the GEO database (West et al., 2014). Nucleosome density was calculated by averaging the MNase sequencing signal over the CGR interval. DNA methylation levels were acquired from published datasets for the cell types in Figure 3 (Bernstein et al., 2010; dos Santos et al., 2015; Guo et al., 2017; Hon et al., 2013; Kunde-Ramamoorthy et al., 2014; Lister et al., 2009; Lu et al., 2014; Vincent et al., 2013; Ziller et al., 2013). DNA methylation scores were obtained by averaging all individual CpG methylation scores within the CGR. The hg19 reference genome was used for this analysis because all of the published datasets were mapped in hg19.

BAC Modification and Preparation

The human gene-desert RP11-722D BAC was purchased from CHORI-BACPAC. Exogenous sequences were inserted into the BAC as described (Gong and Yang, 2005). BACs were electroporated into SW102 RecA-expressing bacteria and selected for targeted recombination of GalK and replacement of GalK by minimal galactose medium or deoxygalactose, respectively (Warming et al., 2005). A PGK-neomycin-expressing cassette was introduced into the BAC as described (Wang et al., 2001). Successful recombineering was confirmed by restriction enzyme fingerprinting and sequencing of the insert region.

BAC DNA was isolated using the Large Construct Kit (Qiagen) and linearized with the restriction enzyme PI-SceI (New England Biolabs). BACs were pre-methylated as described (Xu et al., 2009) and BAC integrity was verified on a large pulsed-field gel (Bio-Rad CHEF Mapper XA).

Cell Culture

The R1 male mouse ESC line was grown in Knockout DMEM supplemented with 15% fetal bovine serum (Omega), 0.1 mM non-essential amino acids, 2 mM L-glutamine, 1% penicillin/streptomycin, 0.05 mM β-mercaptoethanol, and 1,000 U/mL LIF (ESGRO, Millipore). All culture products were purchased from Gibco unless otherwise noted. ESCs were maintained in gelatin (STEMCELL Technologies)-coated Petri dishes and on a layer of mouse embryonic fibroblasts mitotically inactivated with mitomycin C. ESCs were removed from plates using trypsin-EDTA (STEMCELL Technologies).

ESCs were grown to confluency on a 10-cm plate prior to transduction with 5–20 μg of BAC DNA by electroporation at 0.27 kV, 500 μF. After a short recovery, the ESCs were replated 1:2. G418/neomycin at 255 μg/μL was added for approximately 10 days to select transfected clones. Single colonies were picked and expanded in G418. Genomic DNA was isolated from stable ESC clones with the DNeasy kit (Qiagen). Integration of BAC DNA was confirmed by genotyping PCR.

Bisulfite Sequencing and ChIP

Bisulfite treatment of 2.5 μg of genomic DNA was performed as described (Millar et al., 2002). Sequence-specific PCR of the bisulfite-treated DNA was performed using primers specific to BAC regions. The PCR fragments were cloned into the pCRII vector (Invitrogen, K2070-20) and transformed into DH5α *E. coli* cells. Miniprep plasmid DNA was sequenced using M13 reverse primers (5′-AGGAAACAGCTATGACCAT-3′).

Nuclei from approximately 30 million ESCs were isolated as previously described (Ramirez-Carrozzi et al., 2006). The nuclei were then supplemented with protease inhibitors (α-Complete, Roche) and sonicated in a Diagenode Bioruptor Twin sonicator for 15 min with 30-s cycles. One hundred micrograms of chromatin was incubated overnight at 4°C with 5 μg of an antibody to H3K4me3 (Millipore, 07-473). Chromatin complexes were recovered by binding to Protein A Dynabeads (Invitrogen, 100-02D) and were released from the beads by elution with $NaCHO_3$ 1% SDS buffer, and cross-linking was reversed by incubation at 65°C overnight. DNA was purified using the PCR purification kit (Qiagen). The quantity of immunoprecipitated DNA was measured by qPCR on an iCycler (Bio-Rad). The amount of DNA for each primer set was calculated relative to a 5% input chromatin control sample. To control for variable BAC integrants, the percentage inputs for all BAC regions were normalized to the percentage input at a downstream BAC CpG island with consistent enrichment.

Data and Code Availability

The code used to find CGRs is available at https://github.com/teneth/findCGR/.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.stemcr.2021.02.016.

AUTHOR CONTRIBUTIONS

J.L. and S.T.S. designed the experiments; J.L. performed the experiments; J.L., D.L., M.P., and S.T.S. analyzed the data; and J.L. and S.T.S wrote the manuscript.

CONFLICTS OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank J. Ernst, K. Plath, A. Smith, and K. Zaret, for helpful discussions and/or critical reading of the manuscript, and L. Smale for preparation of the graphical abstract. This work was supported by

NIH grants R01 GM086372 (S.T.S.), P01 GM099134 (S.T.S.), and T32 GM007185 (J.L.).

Received: July 1, 2020
Revised: February 22, 2021
Accepted: February 23, 2021
Published: March 25, 2021

REFERENCES

Agarwal, V., and Shendure, J. (2020). Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. *31*, 107663.

Auclair, G., Guibert, S., Bender, A., and Weber, M. (2014). Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. Genome Biol. *15*, 545.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature *520*, 243–247.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome Prediction. PLoS Comput. Biol. *3*, e110.

Cierpicki, T., Risner, L.E., Grembecka, J., Lukasik, S.M., Popovic, R., Omonkowska, M., Shultis, D.D., Zeleznik-Le, N.J., and Bushweller, J.H. (2010). Structure of the MLL CXXC domain–DNA complex and its functional role in MLL-AF9 leukemia. Nat. Struct. Mol. Biol. *17*, 62–68.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.-H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev. *26*, 1714–1728.

Cohen, N.M., Kenigsberg, E., and Tanay, A. (2011). Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell *145*, 773–786.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

dos Santos, C.O., Dolzhenko, E., Hodges, E., Smith, A.D., and Hannon, G.J. (2015). An epigenetic memory of pregnancy in the mouse mammary gland. Cell Rep. *11*, 1102–1109.

Drew, H.R., and Travers, A.A. (1985). DNA bending and its relation to nucleosome positioning. J. Mol. Biol. *186*, 773–790.

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

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

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

Gong, S., and Yang, X.W. (2005). Modification of bacterial artificial chromosomes (BACs) and preparation of intact BAC DNA for generation of transgenic mice. Curr. Protoc. Neurosci. *31*, 5.21.1–5.21.13.

Guo, G., von Meyenn, F., Rostovskaya, M., Clarke, J., Dietmann, S., Baker, D., Sahakyan, A., Myers, S., Bertone, P., Reik, W., et al. (2017). Epigenetic resetting of human pluripotency. Development *144*, 2748–2763.

Hartl, D., Krebs, A.R., Grand, R.S., Baubec, T., Isbel, L., Wirbelauer, C., Burger, L., and Schübeler, D. (2019). CG dinucleotides enhance promoter activity independent of DNA methylation. Genome Res. *29*, 554–563.

Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., and Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat. Genet. *45*, 1198–1206.

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

Krebs, A.R., Dessus-Babus, S., Burger, L., and Schübeler, D. (2014). High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. ELife *3*, e04094.

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

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

Lövkvist, C., Dodd, I.B., Sneppen, K., and Haerter, J.O. (2016). DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. *44*, 5123–5132.

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

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S., and Zhang, Y. (2014). Role of Tet proteins in enhancer activity and telomere elongation. Genes Dev. *28*, 2103–2119.

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

Millar, D.S., Warnecke, P.M., Melki, J.R., and Clark, S.J. (2002). Methylation sequencing from limiting DNA: embryonic, fixed, and microdissected cells. Methods *27*, 108–113.

Morrison, A.J., and Shen, X. (2009). Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. Nat. Rev. Mol. Cell Biol. *10*, 373–384.

Morselli, M., Pastor, W.A., Montanini, B., Nee, K., Ferrari, R., Fu, K., Bonora, G., Rubbi, L., Clark, A.T., Ottonello, S., et al. (2015). In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. ELife *4*, e06205.

Nichols, J., and Smith, A. (2009). Naive and primed pluripotent states. Cell Stem Cell *4*, 487–492.

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

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

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell *158*, 1254–1269.

The Mouse ENCODE Consortium, Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic identification of culture conditions for induction and maintenance of naïve human pluripotency. Cell Stem Cell *15*, 524–526.

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

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science *346*, 1007–1012.

Vincent, J.J., Huang, Y., Chen, P.-Y., Feng, S., Calvopiña, J.H., Nee, K., Lee, S.A., Le, T., Yoon, A.J., Faull, K., et al. (2013). Stage-specific roles for Tet1 and Tet2 in DNA demethylation in primordial germ cells. Cell Stem Cell *12*, 470–478.

Wang, Z., Engler, P., Longacre, A., and Storb, U. (2001). An efficient method for high-fidelity BAC/PAC retrofitting with a selectable marker for mammalian cell transfection. Genome Res. *11*, 137–142.

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

West, J.A., Cook, A., Alver, B.H., Stadtfeld, M., Deaton, A.M., Hochedlinger, K., Park, P.J., Tolstorukov, M.Y., and Kingston, R.E. (2014). Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. Nat. Commun. *5*, 4719.

Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. Biostatistics *11*, 499–514.

Xu, C., Liu, K., Lei, M., Yang, A., Li, Y., Hughes, T.R., and Min, J. (2018). DNA sequence recognition of human CXXC domains and their structural determinants. Structure *26*, 85–95.

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

Yu, N., Guo, X., Zelikovsky, A., and Pan, Y. (2017). GaussianCpG: a Gaussian model for detection of CpG island in human genome sequences. BMC Genomics *18* (*Suppl 4*), 392.

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

Stem Cell Reports, Volume *16*

Supplemental Information

Species-Specific Relationships between DNA and Chromatin Properties of CpG Islands in Embryonic Stem Cells and Differentiated Cells

Justin Langerman, David Lopez, Matteo Pellegrini, and Stephen T. Smale

Supplemental Figures for:

Species-Specific Relationships between the DNA and Chromatin Properties of CpG Islands in Embryonic Stem Cells and Differentiated Cells

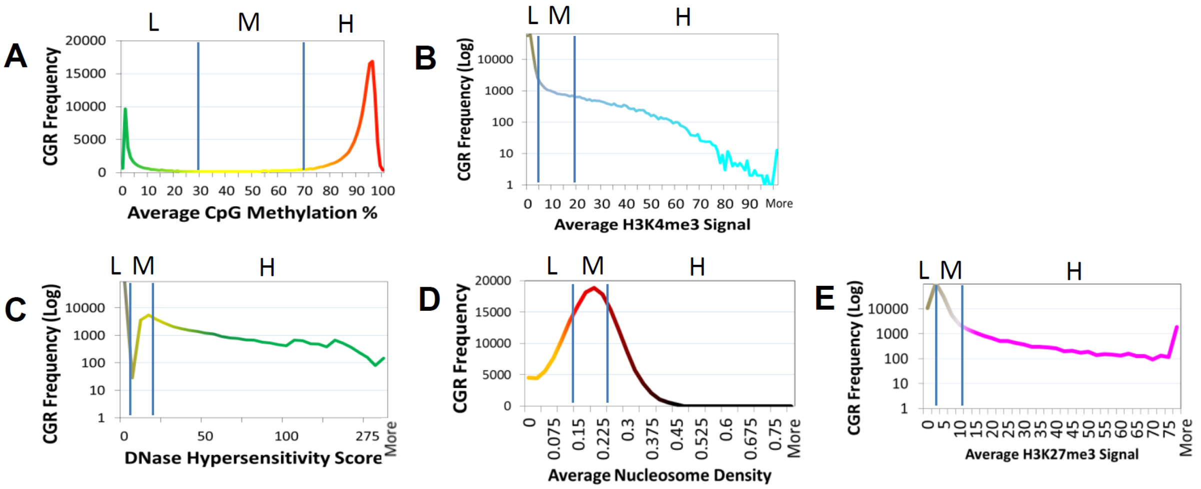


Figure S1. Distributions of chromatin properties at human ESC CGRs. Related to Figure 2.
Shown are graphs of the chromatin scores for all CGRs in human ESC. Lines on each graph demarcate the ranges used for low, medium, and high chromatin bins in subsequent figures, which are labeled with an L, M, or H respectively. Each line shows the count of CGRs at each integer value for chromatin scores pertaining to (A) average DNA methylation across the CGR, (B) average H3K4me3 ChIP-Seq signal over the CGR, (C) peak DNase HS score, (D) average MNase-seq nucleosome density, or (E) average H3K27me3 ChIP-Seq signal, in human ESC.

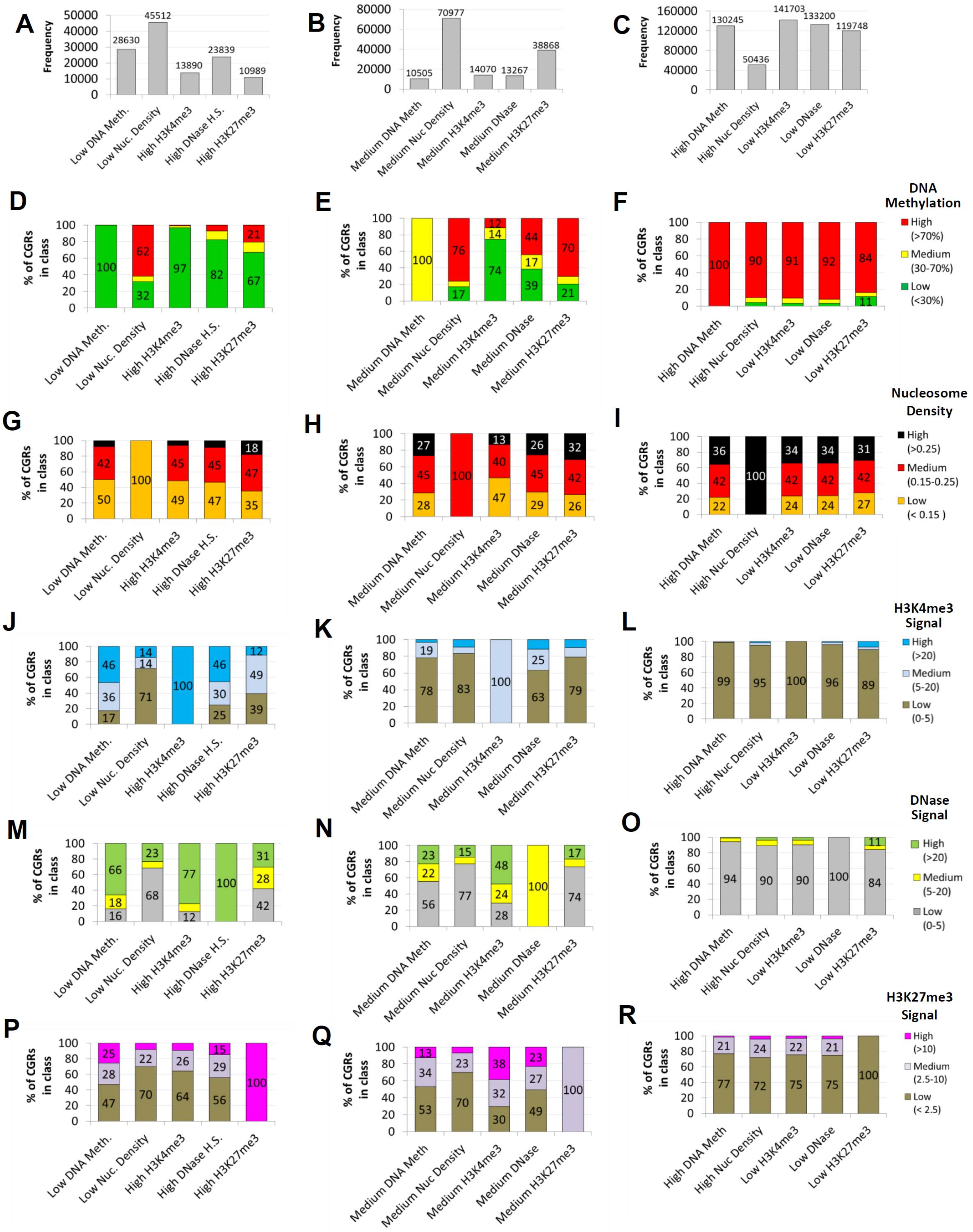


Figure S2. Intercorrelation of chromatin features at human ESC CGRs. Related to Figure 2.
To examine relationships between the chromatin features, co-segregation analyses were performed after assigning each human ESC CGR to one of three bins for each feature. The three bins for each chromatin feature are displayed in Figure S1. Bins for each chromatin feature were defined somewhat arbitrarily, but with the goal of creating three bins of approximately equal size, while looking for natural cutoffs based on the distributions shown in Figure S1.

(A-C) The numbers of CGRs (173,307 total CGRs) within each low, medium, or high bin for each chromatin feature are shown (see L, M, and H in Figure S1).
(D-R) The co-segregation analysis of chromatin features is shown. This analysis revealed a high degree of co-segregation for CGRs with low CpG methylation, high or medium H3K4me3, and high or medium DNase HS. For example, 97% of CGRs with high H3K4me3 exhibited low DNA methylation (panel D, bar 3; n=13,890 CGRs with high H3K4me3, as shown in panel A. In contrast, 91% of CGRs with low H3K4me3 exhibited high DNA methylation (panel F, bar 3; n=141,703, as shown in panel C) .
Strong correlations between CpG methylation, H3K4me3, and DNase HS were also observed when focusing attention on H3K4me3 levels (J-L) or DNase HS levels (M-O). Interestingly, nucleosome density displayed a relatively weak correlation to these three chromatin features. Although 90% of CGRs with high nucleosome density exhibited high DNA methylation (F), only 32% of CpGs with low nucleosome density exhibited low DNA methylation (D). Similarly, although 90% of CGRs with high nucleosome density exhibited low DNase HS (O), only 23% of CGRs with low nucleosome density exhibited high DNase HS (M). To be sure, CGRs with low DNA methylation, high H3K4me3, and high DNase HS exhibited moderately reduced nucleosome density distributions in comparison to CGRs with the opposite features (compare G to I). However, this relatively weak correlation between nucleosome density and the other chromatin features can be accounted for by differences in genome location. Importantly, the weak correlation between low nucleosome density and the other chromatin features could not be strengthened by using different bin thresholds (data not shown). Histone H3K27me3 also co-segregated only weakly with the other four chromatin properties (D-R), consistent with current knowledge that H3K27me3 is deposited at only a subset of inactive islands. These weak correlations are most readily apparent in panels P, Q, and R, which show that most CGRs exhibit low H3K27me3 levels, regardless of the other chromatin properties observed at the CGR.

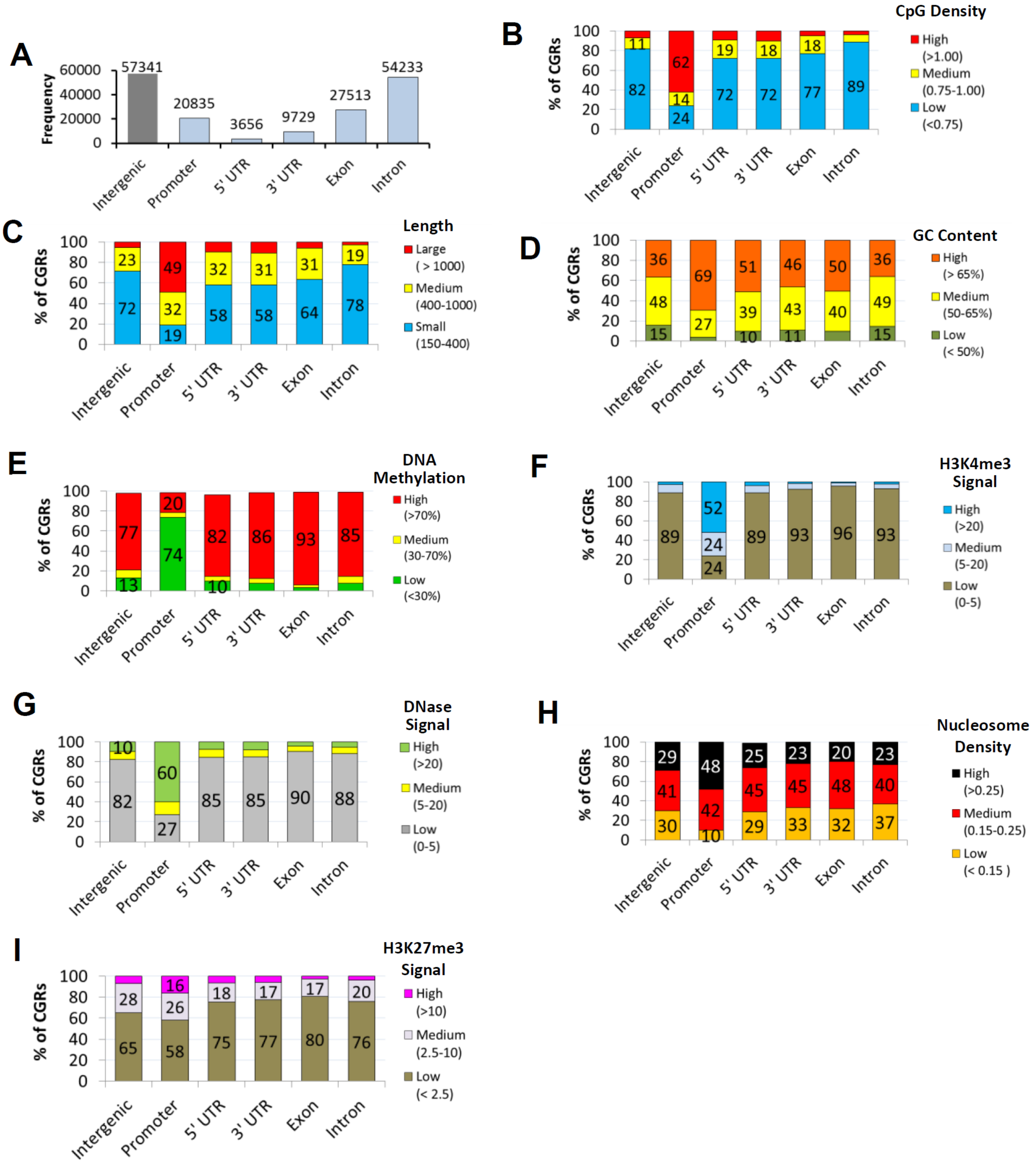


Figure S3. **Features of human ESC CGRs are influenced by genomic location. Related to Figure 2.** The interrelationship between DNA properties and chromatin properties is shown. Importantly, we found that this analysis benefited greatly from simultaneous consideration of the genomic locations of the CGRs.

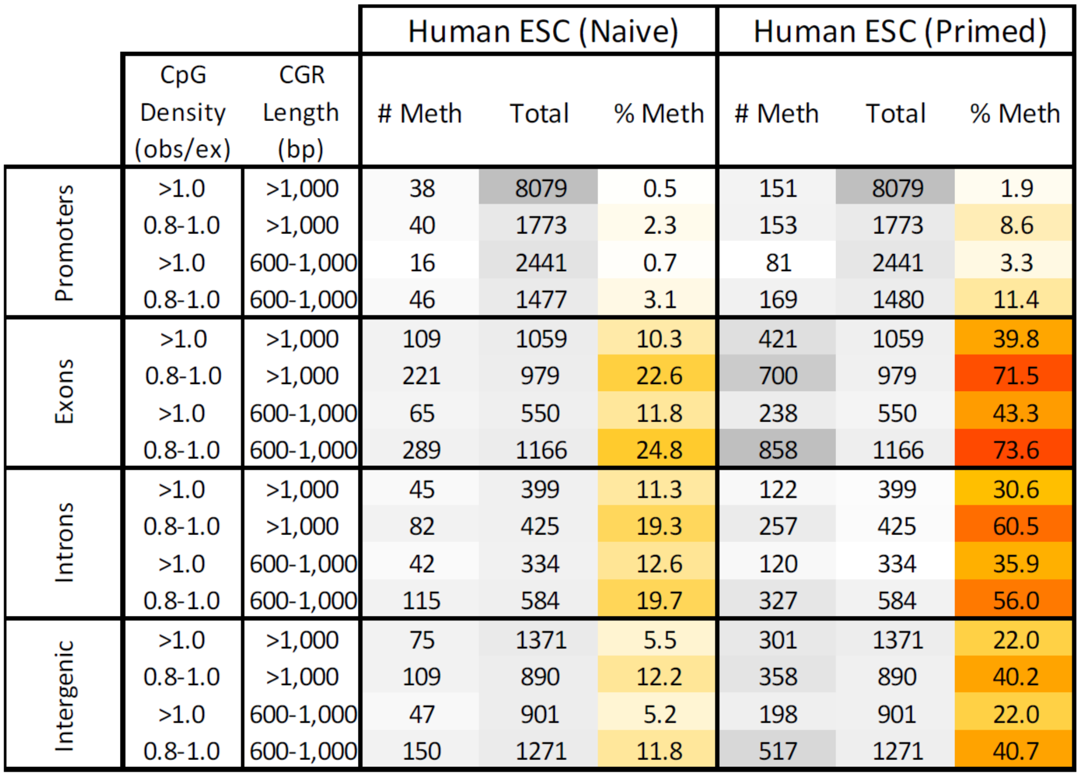
(A) 12% of all CGRs are located in promoter regions, with the remaining 88% distributed among 5' and 3' untranslated regions (UTRs), exons, introns, and intergenic regions. Multiple location matches are superseded in the following order: Promoter, UTR, Exon, and then Intron.

(B) After assigning CGRs to bins on the basis of defined DNA properties, we found that promoter CGRs exhibit a higher CpG density distribution than non-promoter CGRs.

(C-D) Promoter CGRs are also generally longer and possess higher GC percentages than non-promoter CGRs.

(E-H) Promoter CGRs often exhibited lower DNA methylation levels, higher H3K4me3, higher DNase HS, and higher nucleosome densities in human ESC than non-promoter CGRs.

(I) A slightly larger fraction of promoter CGRs exhibited higher H3K27me3 levels than non-promoter CGRs in human ESC.



			Human ESC (Naive)			Human ESC (Primed)		
	CpG Density (obs/ex)	CGR Length (bp)	# Meth	Total	% Meth	# Meth	Total	% Meth
Promoters	>1.0	>1,000	38	8079	0.5	151	8079	1.9
	0.8-1.0	>1,000	40	1773	2.3	153	1773	8.6
	>1.0	600-1,000	16	2441	0.7	81	2441	3.3
	0.8-1.0	600-1,000	46	1477	3.1	169	1480	11.4
Exons	>1.0	>1,000	109	1059	10.3	421	1059	39.8
	0.8-1.0	>1,000	221	979	22.6	700	979	71.5
	>1.0	600-1,000	65	550	11.8	238	550	43.3
	0.8-1.0	600-1,000	289	1166	24.8	858	1166	73.6
Introns	>1.0	>1,000	45	399	11.3	122	399	30.6
	0.8-1.0	>1,000	82	425	19.3	257	425	60.5
	>1.0	600-1,000	42	334	12.6	120	334	35.9
	0.8-1.0	600-1,000	115	584	19.7	327	584	56.0
Intergenic	>1.0	>1,000	75	1371	5.5	301	1371	22.0
	0.8-1.0	>1,000	109	890	12.2	358	890	40.2
	>1.0	600-1,000	47	901	5.2	198	901	22.0
	0.8-1.0	600-1,000	150	1271	11.8	517	1271	40.7

Figure S4. **Methylation of human CGRs differs dramatically between the naïve and primed ESC states even at strict criteria. Related to Figure 3.**
As in Figure 3, the tables show the counts of CGRs in human DNA methylation data sets that meet the DNA property criteria labeled at left. CGRs are separated by genomic location and by non-overlapping CpG density and length ranges, for human ESC naïve and primed condition methylomes (Guo et al., 2017). The number of CGRs with high DNA methylation (>70%) in each criteria range is shown under "# Meth" next to the number of regions that qualify for the criteria, "Total". The last column for each methylome group is the percent of CGRs methylated for each criterion, "% Meth". The tables are colored by frequency; increased grey indicates higher CGR numbers, while increased red indicates a higher percentage in the "% Meth" column. CGRs with insufficient bisulfite sequencing reads were discarded.