Cell

Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates

Antoine Molaro,[1,3] Emily Hodges,[1,3] Fang Fang,[2] Qiang Song,[2] W. Richard McCombie,[1] Gregory J. Hannon,[1,*] and Andrew D. Smith[2,*]
[1]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[2]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[3]These authors contributed equally to this work
*Correspondence: hannon@cshl.edu (G.J.H.), andrewds@usc.edu (A.D.S.)

DOI 10.1016/j.cell.2011.08.016

SUMMARY

During germ cell and preimplantation development, mammalian cells undergo nearly complete reprogramming of DNA methylation patterns. We profiled the methylomes of human and chimp sperm as a basis for comparison to methylation patterns of ESCs. Although the majority of promoters escape methylation in both ESCs and sperm, the corresponding hypomethylated regions show substantial structural differences. Repeat elements are heavily methylated in both germ and somatic cells; however, retrotransposons from several subfamilies evade methylation more effectively during male germ cell development, whereas other subfamilies show the opposite trend. Comparing methylomes of human and chimp sperm revealed a subset of differentially methylated promoters and strikingly divergent methylation in retrotransposon subfamilies, with an evolutionary impact that is apparent in the underlying genomic sequence. Thus, the features that determine DNA methylation patterns differ between male germ cells and somatic cells, and elements of these features have diverged between humans and chimpanzees.

INTRODUCTION

In mammals, proper DNA methylation is essential for both fertility and viability of offspring (Bestor, 1998; Bourc'his and Bestor, 2004; Li et al., 1992; Okano et al., 1999; Walsh et al., 1998). DNA methylation in germ cells is required for successful meiosis (Bourc'his and Bestor, 2004), and blastocysts derived from embryonic stem cells (ESCs) lacking DNA methyltransferases (DNMTs) cannot survive past approximately 10 days of development (Li et al., 1992).

Mammalian germ cells are derived from somatic cells, rather than being set-aside during the first zygotic cleavages. During germ cell development, the genome undergoes a wave of nearly complete demethylation and remethylation (Popp et al., 2010; Walsh et al., 1998). This reprogramming event correlates with re-establishment of totipotency and with the creation of sex-specific methylation patterns at imprinted loci (reviewed by Sasaki and Matsui, 2008). Germ cell methylation patterns are erased and reset during a second wave of epigenetic reprogramming that occurs during preimplantation development. Post-fertilization, DNA methylation levels reach a nadir around the eight-cell stage, after which methylation is rewritten, attaining its somatic level by the blastocyst stage (Mayer et al., 2000). Because this is completed prior to the establishment of the inner cell mass from which cultured ESCs are derived, one can view ESCs and mature germ cells as the terminal products of the two landmark epigenetic reprogramming events in mammals.

Mobile genetic elements constitute roughly half of most mammalian genomes (Lander et al., 2001). Repression of transposons relies critically on DNA methylation and is essential for the maintenance of genomic stability in the long term and of germ cell function in the near term (Bestor, 1998; Bourc'his and Bestor, 2004; Okano et al., 1999; Walsh et al., 1998). At least in part, silencing of repeated DNA depends upon an abundant class of PIWI-associated small RNAs, called piRNAs (reviewed in Aravin and Hannon, 2008). In the absence of this pathway, methylation is lost on at least some element copies, transposons are derepressed, and germ cell development is arrested in meiosis.

CpG dinucleotides are underrepresented in mammalian genomes, most likely because a higher rate of spontaneous deamination of methylated cytosines exerts evolutionary pressure for CpG depletion by frequent CpG-to-TpG transitions (Duncan and Miller, 1980; Ehrlich et al., 1990). Mammalian genomes contain areas of relatively high CpG density, called "CpG islands" (CGIs) (Gardiner-Garden and Frommer, 1987), which have avoided CpG depletion over evolutionary time. CGIs are frequently observed at promoters and in some cases have been shown to exert regulatory effects. Thus, selection against CpG depletion may reflect the importance of specific CpG dinucleotides as sequence-based binding sites or simply the requirement for a certain regional density of CpGs. As an alternative, the existence of CGIs may simply be an artifact of longstanding hypomethylation of these regions, and consequent

Table 1. Shotgun Bisulfite Sequencing of Human and Chimp Sperm Methylomes

Species	Sample	Mapped	Distinct	Mismatches	BS Conversion	Methylation	CpG Coverage	CpGs Covered
Human	sperm (1)	609,127,589	388,835,058	1.58	0.992	0.724	8.8	0.96
	sperm (2)	588,920,777	316,860,245	1.84	0.983	0.674	7.3	0.94
	sperm (both)	1,198,048,366	705,695,303	1.70	0.988	0.701	16.1	0.96
	ESCs	940,731,922	366,844,212	0.64	0.988	0.663	14.1	0.93
Chimp	sperm (1)	459,258,834	255,193,493	1.87	0.985	0.665	6.2	0.95
	sperm (2)	520,905,232	327,796,614	1.70	0.984	0.672	7.4	0.94
	sperm (both)	980,164,066	582,990,107	1.78	0.985	0.669	13.6	0.96

Mapped: reads mapping optimally to a single location in the reference genome. Distinct: number of genomic locations to which a read maps; when multiple reads map to the same position, one with the best mapping score was selected at random, and all others discarded. Mismatches: average number of mismatches for the reads indicated in the distinct fragments column. Bisulfite (BS) conversion rate was calculated at non-CpG cytosines. Methylation: proportion of Cs in reads mapping over CpG dinucleotides.

relief from CpG erosion, in mammalian germ cells. Under this hypo-deamination model, selective pressure is independent of CpG density, per se, and CGIs may instead be a secondary consequence of protection from methylation at specific sites combined with prevalent methylation elsewhere in the genome (Cooper and Krawczak, 1989; Duncan and Miller, 1980; Ehrlich et al., 1990).

Studies encompassing evolutionarily distant species have shown that broad features of the epigenome, such as the high methylation levels of gene bodies and repeats, are deeply conserved (Zemach et al., 2010). In closely related species, however, fine-scale analysis of DNA methylation state reveals variation. The chimpanzee and human genomes share more than 95% sequence homology but display regions of differential methylation (Enard et al., 2004). Through focused studies, we have gained glimpses into the characteristics of the methylome and the evolutionary pressures that shape it. We wished to enable genome-wide comparisons of DNA methylation states in closely related species and to examine possible differences between the two major waves of epigenetic remodeling that occur during the mammalian life cycle. We therefore produced full-genome, single-CpG resolution DNA methylation profiles in human and chimp sperm and compared these with methylation maps from human ESCs (Laurent et al., 2010).

RESULTS

Methylomes of Mature Male Germ Cells in Human and Chimp

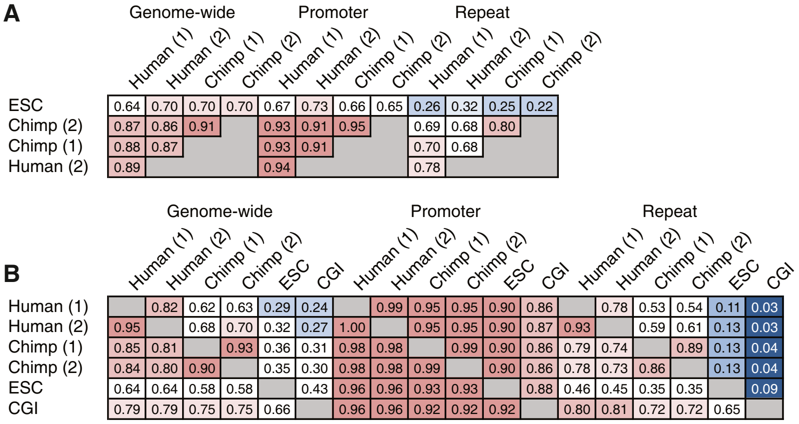
We conducted genome-wide shotgun bisulfite sequencing of sperm DNA samples isolated from two human and chimp donors (see Extended Experimental Procedures for details). Basic data analysis was conducted using a custom pipeline. We were able to determine methylation status for 96% of genomic CpGs in the human and chimp samples from a total of 28 million and 27 million CpGs, respectively (Table 1). Read coverage for CpGs on autosomes averaged 16× in human with an overall methylation level of ~70% for all CpG sites. For chimp we sequenced to an average coverage of nearly 14× and observed an average methylation level of ~67%. We did not observe significant methylation at non-CpG sites in either dataset. For comparison, we applied our analysis pipeline to a whole-genome bisulfite dataset from human ESCs (Laurent et al., 2010). This dataset was comparable to our own, with 93% of CpG dinucleotides covered and an average depth of 14× on CpGs genome-wide.

We identified contiguous domains of low methylation, termed hypomethylated regions or HMRs, in a manner independent of genomic annotations such as CGIs and promoters. Because methylation levels in sperm were generally high, HMRs appeared obvious on browser plots as valleys in which methylation dropped to very low levels. To call HMRs in a statistically principled manner, we designed a novel computational approach, based on a two-state hidden Markov model with Beta-Binomial emission distributions (see Extended Experimental Procedures). This algorithm identified ~79k HMRs in human sperm and ~70k HMRs in chimp sperm. Only ~44.5k HMRs were identified using the human ESC dataset, despite similar sequence coverage and overall methylation level (Laurent et al., 2010; see Table 1 and Table S1A available online). The sizes of HMRs also differed between germ and ESCs. In both chimp and human sperm, the mean size of HMRs was ~1.8 kb, and the median was ~1.3 kb. In ESCs, HMRs showed a mean size of ~1.2 kb with a median of 833 bp. HMRs overlapped all classes of genomic annotation (see Table S1B).

Global Comparisons among Primate Sperm Methylomes and with Human ESCs

Average methylation levels differed by a small amount among the human donors (donor 1: 72%; donor 2: 67%) but were more similar among chimp donors (donors 1 and 2: 67%). The methylation status of individual CpGs of HMRs correlated very highly between individuals, with divergence being higher in repeats as compared to promoters (Figures 1A and 1B). High interindividual correlations at the CpG and the HMR levels imply that our datasets permit accurate calling of CpG methylation genome-wide.

We also compared methylation between species at an individual nucleotide level (see Extended Experimental Procedures for details). As expected, the correlations between human and chimp sperm methylation are high, but the correlation remains generally highest within species.



A

	Genome-wide				Promoter				Repeat			
	Human (1)	Human (2)	Chimp (1)	Chimp (2)	Human (1)	Human (2)	Chimp (1)	Chimp (2)	Human (1)	Human (2)	Chimp (1)	Chimp (2)
ESC	0.64	0.70	0.70	0.70	0.67	0.73	0.66	0.65	0.26	0.32	0.25	0.22
Chimp (2)	0.87	0.86	0.91		0.93	0.91	0.95		0.69	0.68	0.80	
Chimp (1)	0.88	0.87			0.93	0.91			0.70	0.68		
Human (2)	0.89				0.94				0.78			

B

	Genome-wide						Promoter						Repeat					
	Human (1)	Human (2)	Chimp (1)	Chimp (2)	ESC	CGI	Human (1)	Human (2)	Chimp (1)	Chimp (2)	ESC	CGI	Human (1)	Human (2)	Chimp (1)	Chimp (2)	ESC	CGI
Human (1)		0.82	0.62	0.63	0.29	0.24		0.99	0.95	0.95	0.90	0.86		0.78	0.53	0.54	0.11	0.03
Human (2)	0.95		0.68	0.70	0.32	0.27	1.00		0.95	0.95	0.90	0.87	0.93		0.59	0.61	0.13	0.03
Chimp (1)	0.85	0.81		0.93	0.36	0.31	0.98	0.98		0.99	0.90	0.86	0.79	0.74		0.89	0.13	0.04
Chimp (2)	0.84	0.80	0.90		0.35	0.30	0.98	0.98	0.99		0.90	0.86	0.78	0.73	0.86		0.13	0.04
ESC	0.64	0.64	0.58	0.58		0.43	0.96	0.96	0.93	0.93		0.88	0.46	0.45	0.35	0.35		0.09
CGI	0.79	0.79	0.75	0.75	0.66		0.96	0.96	0.92	0.92	0.92		0.80	0.81	0.72	0.72	0.65	

Figure 1. A Global View of Sperm and ESC Methylomes

(A) Correlations between methylomes with methylation levels measured at individual CpG sites. Correlations are displayed for CpGs genome-wide, within promoters, and within repeats, and correlation coefficients are colored blue to red to indicate low to high, respectively.

(B) Overlap between sets of HMRs from human sperm, chimp sperm, and ESC methylomes, along with annotated CGIs. Each cell gives the fraction of HMRs corresponding to the row that overlaps HMRs corresponding to the column. Colors are overlaid as in (A).

See also Table S1.

We also directly compared the methylomes from each of the human and chimp donors with the human ESC methylome. The nucleotide-level correlations between sperm methylation of each of the four primate individuals were higher than their correlations with ESC methylation patterns (Figure 1A). However, the human ESC methylome did show substantially higher correlation with the human germ cell methylomes than with those of chimp donors. Considered together these results indicate that, although waves of reprogramming in developing germ cells and embryos culminate in high genome-wide methylation, these two methylomes bear substantial differences overall.

Comparison of Hypomethylated Promoters between Sperm and ESC Methylomes

The majority of promoters are associated with HMRs in both sperm and ESCs, indicating widespread bookmarking of promoters during both waves of epigenetic reprogramming. A number of promoters did show differential methylation, with 1336 showing sperm-specific HMRs but only 201 showing ESC-specific HMRs (Figure 2A). Promoters hypomethylated in germ cells were strongly enriched for putative binding sites of transcription factors known to function in testis, including NRF1, NF-Y, YY1, and CREB (see Figure S1). A similar analysis of ESC-specific HMRs failed to yield significant results.

Only the genes with sperm-specific promoter hypomethylation revealed a strong enrichment for functional Gene Ontology (GO) categories. These were associated with germ cell functions (Figure 2B; Table S2) at distinct stages of gametogenesis (e.g., embryonic germ cell development and spermiogenesis). Thus, genes acting at developmental stages, potentially separated by decades, appear to maintain a permissive epigenetic state. Of the eight genes analyzed from the piRNA metabolic process category, seven showed promoter hypomethylation in sperm but not in ESCs, and one was hypomethylated in both (Figure 2B).

Retention of histones in human sperm was reported to be extensive (Hammoud et al., 2009). Our analysis of this data revealed a strong correlation between retained histones marked by H3K4me3 and HMRs at promoters. Among the 25.8k promoters marked by H3K4me3 in sperm, 91% overlapped an identified HMR. In general, these results support prior observations that the presence of H3K4me3 at promoters is often accompanied by hypomethylation (Hammoud et al., 2009; Ooi et al., 2007).

It was previously posited that genes involved in early embryonic development had a distinct chromatin status in sperm, being hypomethylated, histone-retained, enriched in H3K4me3 marks, and thus poised for expression (Hammoud et al., 2009). At least with respect to DNA methylation, we do not detect a preferential link between HMRs in sperm and developmental regulators but instead widespread HMRs. One potential explanation for this perceived discrepancy is that our comparisons involve sperm and ESCs, whereas prior studies used a differentiated cell type to contrast with sperm.

The genes with promoters that lack HMRs in both sperm and ESCs (n = 5,380; Figure 2A) show strong enrichment for G protein-coupled receptors and genes involved in neurological functions (Tables S2C and S2D). The reason why many of these genes, associated with highly specialized cell types, seem to lack promoter HMRs in sperm and ESCs remains obscure.

Shared HMRs Show Distinct Characteristics in Sperm and ESCs

Differences in average size and CpG densities suggest that the HMRs emerging after germ cell reprogramming differ qualitatively from those emerging after zygotic reprogramming (Figure 3A; Table S1A). The majority of HMRs have CpG density between 1% and 10%, and promoter HMRs fall almost exclusively in this range for the sperm methylomes. Those HMRs falling below 1% CpG density lie almost exclusively in repeats. These are overrepresented in human sperm relative to chimp sperm and human ESCs. Promoter-associated HMRs have sizes concentrated between 1 kb and 10 kb in human and chimp sperm, with an overall trend to be broader than promoter-associated HMRs in ESCs (Figure 3A). A notable increase in CpG density accompanies narrowing of HMRs and results in a significant portion of ESC HMRs with a CpG density above 10%.

To probe structural differences among HMRs in ESCs and sperm, we plotted the average methylation around HMR-associated transcriptional start sites (TSSs), genome-wide (Figure 3B, upper). This revealed a general principle, that a core HMR in ESCs, referred to as a nested HMR (Figure 3B, lower), often lies within an extended HMR in sperm. The median size of nested ESC HMRs is 1,498, less than half the median size of 3,109 for

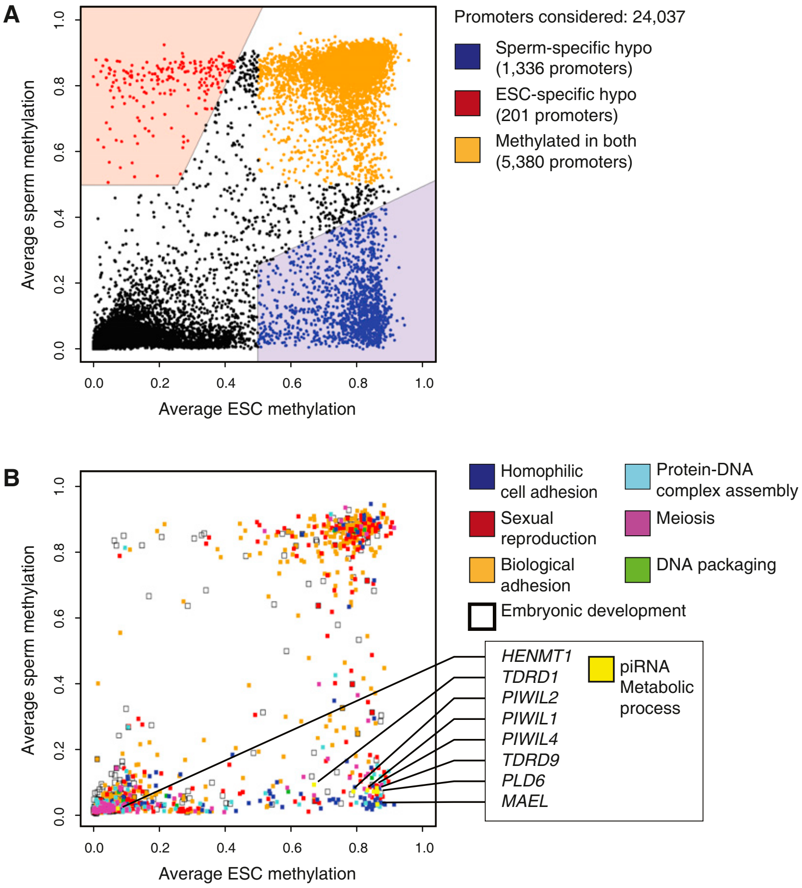


Figure 2. Differentially Reprogrammed Genes and Their Functions

(A) Average methylation through promoters (−1 kbp to +1 kbp) in human sperm and ESCs based on RefSeq gene annotations. Promoters that were hypomethylated only in sperm are shown in blue, those hypomethylated only in ESCs in red, and promoters methylated in both are shaded orange.

(B) Average methylation of promoters associated with GO terms found enriched in the sperm-specific hypomethylated fraction (see A), with the addition of genes from the "embryonic development" term. Individual genes involved in the "piRNA metabolic process" are indicated as an example.

See also Figure S1 and Table S2.

the sperm HMRs in which they reside. This phenomenon was also observed independently in a comparison of somatic and sperm HMRs, where variations in boundaries were additionally correlated with tissue-specific expression (Hodges et al., 2011). Extended HMRs are reminiscent of the concept of CpG shores (Doi et al., 2009), though in comparisons of sperm and ESCs, we made no attempt to correlate gene expression with the widespread phenomenon of nesting that we report herein.

The observation of nested HMRs could arise either from a true expansion of the hypomethylated domain in sperm or as an artifact of sperm having less precise HMR boundaries than ESCs. Examining degrees of change in methylation states across boundary CpGs in both cell types supports the former conclusion (Figure 3C). Thus, nesting appears to represent a general phenomenon and likely reflects differences in the underlying mechanisms by which the boundaries of hypomethylated regions are determined during the waves of de novo methylation that lead to sperm and ESCs.

As a step toward addressing such mechanisms, we asked whether any features are associated with HMR boundaries in either cell type. Two interesting characteristics emerged. Approaching the boundaries of either the extended sperm HMRs or the nested ESC HMRs, CpG densities dropped just prior to the start of the HMR and rose dramatically again thereafter, though overall densities were higher in the nested portions (Figure 3D). This reflects an increase in the average inter-CpG distance at the boundaries of HMRs (Figure 3E). Because our method of identifying HMRs is agnostic to inter-CpG distance, this is not simply an artifact of our approach. One could imagine increases in inter-CpG distance interrupting a processive activity, preventing the spread of de novo methylation either directly or indirectly.

Though we had no a priori expectation that sequence features would reside at sperm or ESC HMR boundaries, we searched for motifs that might occur at or near boundary CpGs, independent of CpG density. We noted a trend toward enrichment for an ACGT motif at ESC boundary CpGs with a corresponding depletion immediately outside ESC HMRs (Figure S2). This pattern was not significantly enriched at the boundaries of extended sperm HMRs. Building upon this observation, we also searched for larger motifs, focusing on those containing a central CpG core. Patterns with strong differences across HMR boundaries tended to have the ACGT core (Table S3). The most enriched pattern for sperm was AACGTT. For ESCs, we saw a well-known E box pattern, CACGTG. Plotting observed-to-expected (o/e) frequencies centered on CpGs around boundaries of extended and nested HMRs (Figure 3F), there was a clear depletion just outside each boundary followed by a sharp enrichment at the boundary CpG for each pattern in the appropriate cell type (Figure S2B). These results raise the possibility that one or more DNA-binding proteins might localize to HMR boundaries during waves of de novo methylation and help to define transitions in methylation states.

Differential Repeat Methylation in Sperm and ESCs

Consistent with prior observations and with the known role of DNA methylation in transposon silencing, most repeat elements were highly methylated in both sperm and ESCs. However, a substantial fraction of HMRs overlapped transposons in chimp and human sperm, with all repeat classes represented (Figure 4A; Table S1B). Fewer repeat-associated HMRs appeared in ESCs. In sperm, HMRs collectively contained 4%–5% of all bases assigned to repeats, compared to 1.3% in ESCs (see Table S1B). Overall, this suggests that different mechanisms,

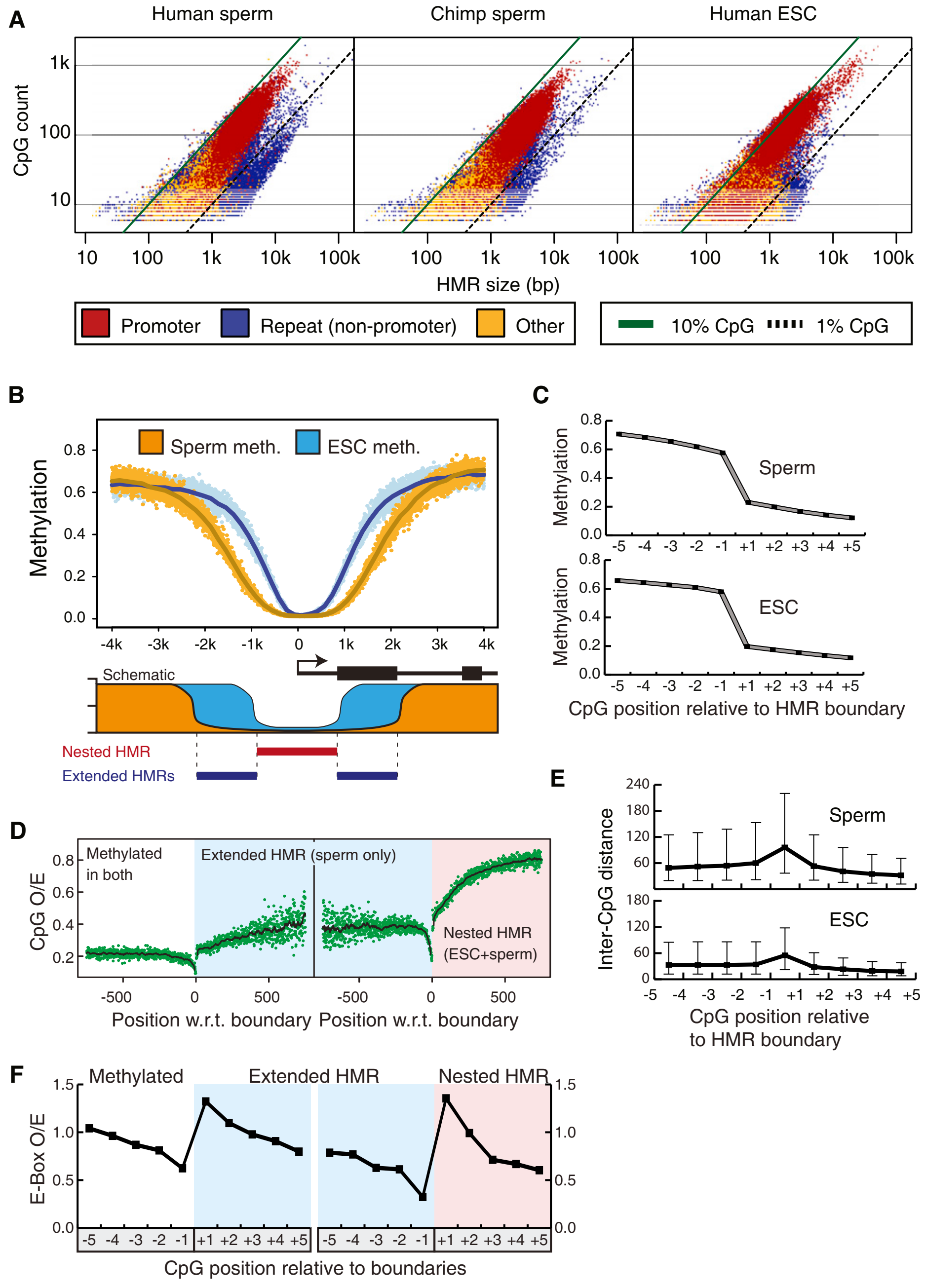


Figure 3. Characteristics of HMRs Emerging from Germline and Somatic Reprogramming

(A) Log-scale plot depicting the sizes (in bases) and numbers of CpGs for all identified HMRs in human sperm (left), chimp sperm (middle), and human ESCs (right). Diagonal lines indicate 10% CpG density (in green) and 1% CpG density (dashed line). HMRs are colored according to promoter overlap (red), overlap with repeats but not promoters (blue), or overlap with neither (orange).

(B) Average methylation around all TSS overlapping HMRs in both sperm (orange) and ESCs (blue); solid lines represent data smoothed using a 20 base sliding window. A schematic depicts the concepts of extended and nested HMRs at promoters.

(C) Average methylation at the −5 to +5 CpGs around boundaries of extended sperm HMRs and nested ESC HMRs (with the +1 CpG defined as the first inside an HMR on either side).

(D) Ratios of observed-to-expected (o/e) CpG density for each nucleotide position relative to boundaries of extended sperm HMRs (left) and nested ESC HMRs (right). Solid lines indicate values smoothed using a 20 base sliding window.

(E) Average inter-CpG distance for −5 to +5 CpGs around HMR boundaries of extended sperm and nested ESC HMRs. Upper and lower quartiles are reported for each position.

(F) Ratio of o/e frequencies of the CACGTG pattern at −5 to +5 CpGs for extended sperm and nested ESC HMRs.

See also Figure S2 and Table S3.

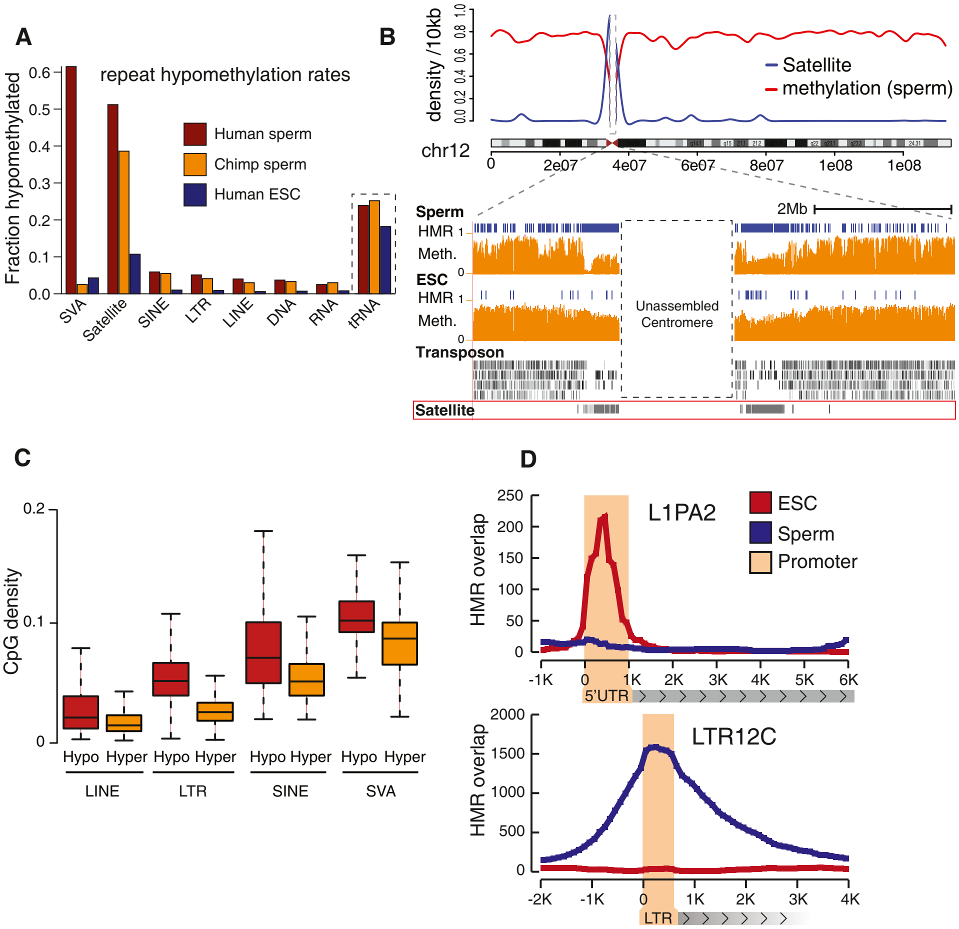


Figure 4. Differential Repeat Methylation during Male Germ Cell and Somatic Reprogramming

(A) For each repeat class, the proportion of elements that overlap HMRs is shown for human sperm (red), chimp sperm (orange), and ESCs (blue).

(B) Upper: Average methylation level (red) and satellite density (blue) in 10 kb sliding windows across chromosome 12. Lower: Chromosome 12 centromeric region with HMRs (blue) and methylation level (orange) for human sperm and ESCs.

(C) CpG densities of hypomethylated repeat copies (red) and methylated repeat copies (yellow) for LINEs, LTRs, SINEs, and SVAs.

(D) HMR overlap distribution around full-length L1PA2 and LTR12 ERV9 elements for human sperm (blue) and ESCs (red).

See also Figure S3 and Table S4.

with different stringencies, direct repeat methylation during germ cell and preimplantation development.

Sperm-Specific Satellite Hypomethylation Is Concentrated at Centromeres

We noted a strong decrease in methylation of sperm DNA within pericentromeric regions, extending several megabases outward from the unassembled core centromeres (Figure 4B). This was not seen in ESCs or in terminally differentiated cells (Hodges et al., 2011). This striking pattern was attributable to sperm-specific hypomethylation of ~75%–80% of the satellite repeats concentrated in pericentromeric regions (Figure 4A). In ESCs, only 16% of pericentromeric satellites were hypomethylated, a figure in accord with the overall hypomethylation rates of nonpericentromeric satellites in ESCs and sperm (Table S4A). Prior studies of mouse germ cells using methylation-sensitive restriction enzymes had noted selectively low methylation at pericentromeric satellites, suggesting that this is a conserved property (Yamagata et al., 2007).

Retroelement Methylation Patterns Are Determined at the Subfamily Level

Proper methylation of retrotransposons is required for transcriptional silencing of full-length and potentially active copies (Bourc'his and Bestor, 2004; Goodier and Kazazian, 2008; Walsh et al., 1998). However, specific retroelements can be active or unmethylated in male germ cells (e.g., AluY and AluYa5) (Schmid, 1991). Given our read lengths, we were able to address the methylation state of virtually all repeat families and most individual copies (see Table S4B).

Overall, retrotransposon copies that were full length or close to consensus showed a slight bias toward hypomethylation (Figures S3A and S3B). However, neither of these attributes could explain the variation observed in retrotransposon

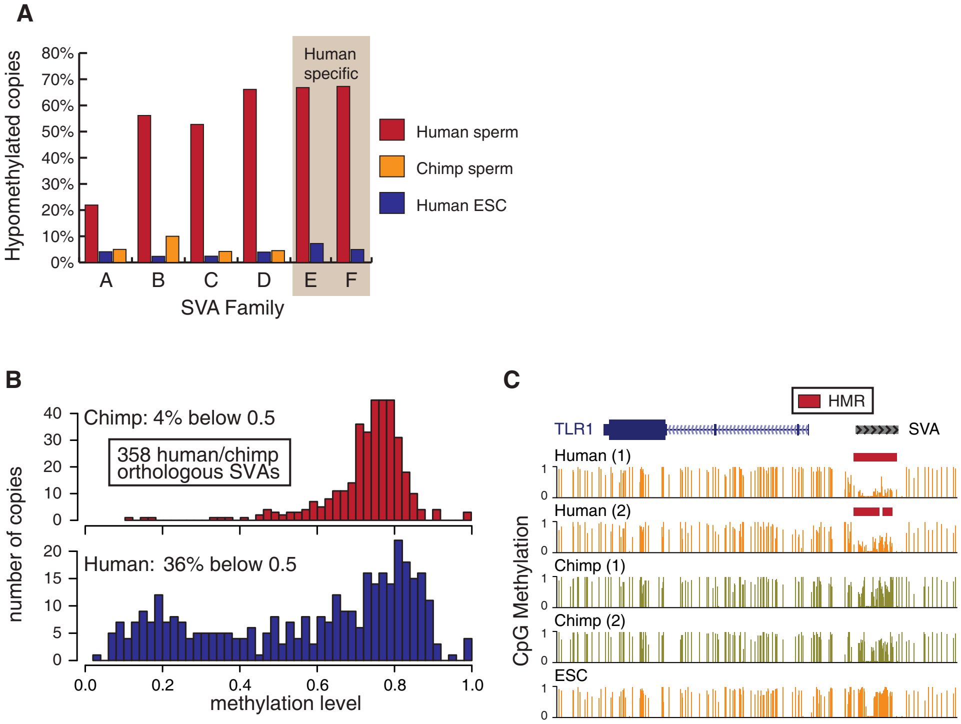


Figure 5. Divergent Methylation of SVA Elements between Human and Chimp

(A) Proportion of hypomethylated SVA copies hypomethylated according to subfamily (A to F) for human sperm (red), chimp sperm (orange), and ESCs (blue).

(B) The distribution of average methylation levels is shown for 358 human (lower) and chimp (upper) SVAs forming high-confidence orthologous pairs.

(C) An SVA insertion shared by human and chimp but with differential methylation between species.

methylation. Hypomethylated repeat copies did tend to have greater CpG density, especially within the LTR and SVA (SINE-R, VNTR, and Alu) classes (Figure 4C). For long interspersed nuclear elements (LINEs), LTR elements, and terminal repeats, HMRs concentrated within regulatory regions, which often show higher CpG density than their coding regions (Figures 4D; Figures S3C and S3D; Tables S4D and S4G). Short interspersed nuclear elements (SINEs) displayed a more uniform hypomethylation (Figure S4E). Thus, similar mechanisms appear to define HMRs in both repeat and nonrepeat portions of the genome, as for most repeats, there is a strong association of sperm HMRs with regulatory regions.

Among the LINEs, subfamilies of L1 were often hypomethylated in both sperm and ESCs, and these trended strongly toward the active groups (Tables S4E and S4H). L1PA subfamilies are considered the most active in the human genome (Khan et al., 2006), and the youngest of these (L1HS and L1PA2) were among the very few subfamilies enriched for hypomethylation in ESCs relative to sperm. Specifically in sperm, we noted hypomethylation of several other L1 families (e.g., L1PA4-16 and L1M3).

Among LTR subfamilies, sperm HMRs were enriched for ERV elements (Table S4C). Hypomethylated copies exist either as part of full-length provirus-like elements or as solo LTRs, with the greatest enrichment for LTRs belonging to "class I" elements (e.g., LTR12; see Tables S4D and S4G). The few LTR subfamilies with more hypomethylated copies in ESCs than sperm are all recently derived, human-specific ERVs (e.g., LTR5 and 13 and HERVH LTR7).

Sperm hypomethylation has been previously reported for primate Alu elements (Kochanek et al., 1993; Liu et al., 1994), and our data revealed several Alu subfamilies with differential methylation in sperm and ESCs, e,g., the AluY subfamily (Tables S4F and S4I). The more precisely defined AluYa5 (human) and AluYd4 (chimp) showed extreme enrichment for hypomethylation in sperm.

Species-Specific Methylation of the SVA Element

SVA elements showed strong, species-specific differences in methylation in human and chimp sperm (Figure 4A). SVAs are composite elements consisting of hexameric repeats, an Alu-like region, a VNTR (variable number of tandem repeats) region, and a SINE-R (Shen et al., 1994). SVA elements were active in the most recent common ancestor of chimp and human (Mills et al., 2006), and multiple examples of neoinsertions suggest that they still cause genomic rearrangements and disease in human (Ostertag et al., 2003).

Among the SVAs, the youngest subfamilies, D–F (Wang et al., 2005), showed the greatest frequency of hypomethylation in human sperm (Figure 5A). Notably, these have a higher CpG density than do older subfamilies. Three hundred and fifty-eight SVA insertions can be assigned as high-confidence orthologs between human and chimp, which remain highly similar in sequence (see Extended Experimental Procedures). Methylation through these element copies was distributed through the full range from very low to very high average methylation, with two modes near 20% and 80% methylation (Figure 5B). In human sperm, 35% of orthologous SVAs had a methylation level below 50%. In sharp contrast, only 6% of copies fell below 50% methylation in chimp. We also annotated 921 SVA elements that appear to represent new insertions occurring after the human-chimp divergence (Mills et al., 2006). 852 (93%) of these were hypomethylated in sperm compared with only 62 (7%) in ESCs (Figure 5A). Considered together, our data indicate that SVA elements have come under different degrees of epigenetic control in the human and chimp lineages.

Many SVA insertions occur at or around promoters (Lander et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005), and these elements often have a CpG content high enough to fit the traditional definition of a CpG island. Given their properties, SVA elements have the potential to introduce differential species- and cell type-specific methylation near genes that may be relevant for their regulation. Figure 5C exemplifies such a situation where, in the case of *TLR1*, no HMR exists near the promoter in chimp sperm or human ESCs, but one is contributed in human sperm by a nearby SVA element. Although sperm are largely transcriptionally silent, similar HMRs are expected to exist in transcriptionally active developing germ cells (data not shown).

Signatures of Selection Accompany Differential Methylation between Primates

CGIs are the most well known evolutionary signature of vertebrate DNA methylation. Their original definition required a CpG o/e ratio of at least 0.6. Although the full set of HMRs in human sperm and ESCs did not reach this empirical cut off, they did pass the 0.4 benchmark used by Weber and colleagues (Figure 6A) (Weber et al., 2007). In general, promoter-associated HMRs did surpass the 0.6 o/e cut off in both sperm and ESCs.

The differences in CpG density in nested and extended HMRs (Figure 3B) imply distinct CpG depletion pressure in these regions. Average CpG composition genome-wide is ~0.2 o/e but reaches ~0.35 in extended HMRs and 0.68 in nested HMRs. We analyzed sperm-specific and ESC-specific HMRs in an attempt to decompose the CpG depletion pressure exerted by the two methylomes. The ESC-specific HMRs reached only 0.35 o/e CpG composition, whereas the sperm-specific HMRs reached a CpG composition of 0.5.

The life cycle of a germ cell can be separated into two components. The first is the time from fertilization to the time that somatically derived primordial germ cells (PGCs) reach the genital ridge. Second is the time during which the PGC develops into a mature germ cell, which contributes to the zygote. The latter period generally spans from birth to the end of the reproductive life of the animal. Our data suggest a model in which methylation patterns present during both of these intervals shape genomic CpG distributions but indicate a greater influence of methylation profiles during germ cell maturation (Figure 6A).

We sought to measure the degree to which differential methylation could lead to CpG decay over the ~6 million years of divergent evolution separating human and chimp. We focused on regions that qualified as HMRs in either chimp or human, as these regions could have either lost methylation along one lineage or gained methylation along the other. For a given regional methylation level, we measured CpG decay as the proportion of regions having lost more than 5% of inferred ancestral CpGs (using gorilla as outgroup) and plotted the relationship between average methylation and decay rate (Figure 6B). The correlation between regional methylation level and CpG decay was extremely strong for both human and chimp. These results indicate that CpG decay is appreciable as a function of methylation even over relatively brief evolutionary periods.

This observation predicted that we might see signatures of selective pressure preventing erosion of some CpGs that are maintained despite germline methylation. To address this question, we analyzed segregating sites at CpG dinucleotides using data from the HapMap 3 project (CEU population; Altshuler et al., 2010). CpGs were treated symmetrically, so each derived allele at these sites can be classified as A, G, or T. As expected, segregating sites with T as the derived allele represent the vast majority.

We generated frequency spectra for each derived allele nucleotide with sites classified according to their methylation level in sperm (Figure S4). As methylation levels increased, derived allele frequencies shifted toward the low ends of the spectra (Figure 6C and Figure S4). This shift was observed not only for derived TpG alleles, which could be explained by an extreme bias in mutation rate, but also for ApG and GpG derived alleles. One interpretation of these findings is that selection is on average weaker at individual CpG sites with lower sperm methylation. Such an interpretation is consistent with recent findings of Cohen et al. (2011), who used sophisticated evolutionary models to posit that selection for high CpG content is not a significant factor contributing to maintenance of CGIs in the genome.

The strong connection between HMRs and gene promoters suggests that the evolutionary gain or loss of HMRs may be associated with changes in selective pressure on functional regulatory regions. To investigate this possibility, we analyzed sequence divergence in HMRs, focusing on those that are human or chimp specific. Because these differentially methylated regions will have different rates of C-to-T transitions, we counted changes from the inferred ancestor only at non-CpG sites. Genomic intervals differing by more than 1% relative to the inferred ancestor were counted as having divergent sequences.

Only 10% of HMRs shared between human and chimp showed divergence from the ancestral sequence at non-CpG sites (Figure 6D). At chimp-specific HMRs, 15% of human sequences and 19% of chimp sequences diverged from the inferred ancestor. At human-specific HMRs, 22% of human sequences diverged and 18% of chimp sequences diverged. These results indicated that changes in methylation state between human and chimp are associated with accelerated non-CpG sequence divergence. Interestingly, in both cases the species with the lower methylation state had a greater rate of divergence, which is consistent with adaptation at novel regulatory regions as a driver for these changes.

We only identified 104 promoters that are hypomethylated in human but not in chimp sperm and only 52 genes with differential promoter methylation in the opposite orientation. Neither set showed significant enrichment for any ontology category.

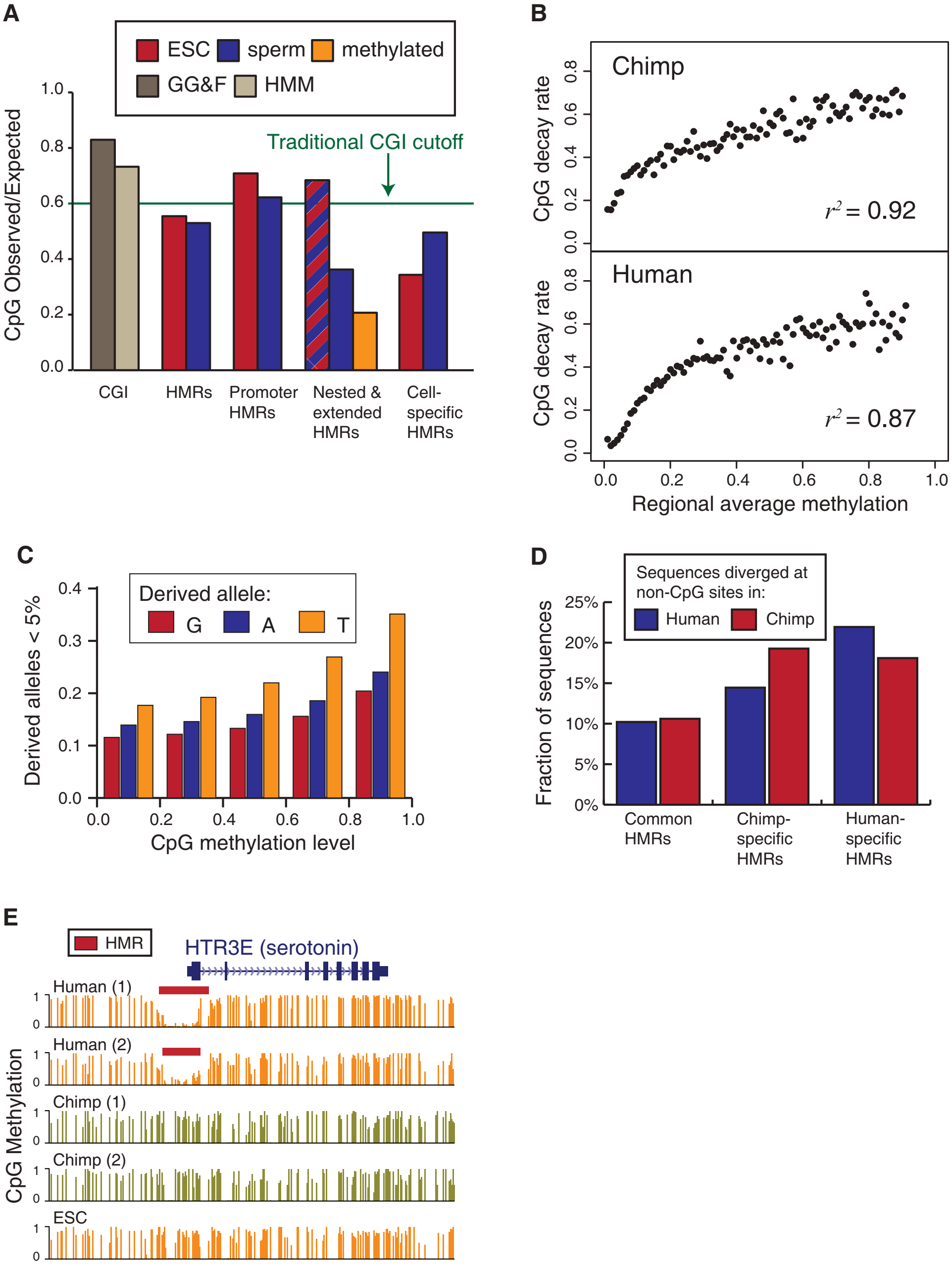


Figure 6. Sequence Features Associated with Methylome Divergence

(A) Ratio of o/e CpG density across all HMRs, those overlapping promoters, those sperm or ESC specific, and the extended/nested HMRs. Data for sperm are indicated in blue and for ESCs are indicated in red; orange indicates ratio immediately outside extended HMRs.

(B) Frequency of regions under CpG decay as a function of methylation for both human and chimp at locations of HMRs in the other species. Decay is presented for chimp in the upper panel and for human in the lower panel.

(C) Frequencies of rare derived alleles at CpG dinucleotides for each derived nucleotide, grouped according to methylation level in human sperm.

(D) Proportion of sequences displaying over 1% nucleotide divergence relative to the inferred ancestor using gorilla as an out-group and counting only non-CpG sites.

(E) The promoter of the human *HTR3E* (serotonin receptor) gene contains an HMR in both human donors but in neither chimp donor.

See also Figure S4 and Table S5.

However, analysis of genes with promoters within 10 kb of an identified human-specific sperm HMR revealed a strong enrichment for neuronal functions (see Table S5). The *HTR3E* gene, a serotonin receptor subunit, is an example of such a gene, whose promoter is selectively hypomethylated in human sperm (Figure 6E).

DISCUSSION

Sperm Methylation Patterns Are Conserved

Overall, sperm methylation patterns were highly similar in all our samples. However, there were differences, even among individuals. There has been much discussion regarding the role of germline transmission of epigenetic marks in interindividual variation (Curley et al., 2011). Changes in epigenetic state could allow flexibility in phenotype that could be reverted over short time spans if a trait became disadvantageous. Erosion of CpG content provides a mechanism to allow fixation of a positive trait in the long run. Thus, changes in DNA methylation patterns preceding changes in DNA sequence presents an attractive model for at least one mode of adaptation. Although evaluating such hypotheses will require many more datasets, the work presented here builds a firm foundation for such studies.

Most Promoters Have HMRs in Sperm

Global resetting of DNA methylation patterns happens twice during mammalian development: once during germ cell development and once early in embryogenesis. Our data permit a genome-scale analysis of these two events. Although high genome-wide levels of methylation are re-established during both waves of epigenetic remodeling, some regions are protected and establish HMR boundaries that appear relevant even in fully differentiated somatic cells (Hodges et al., 2011). A few promoters showed selective hypomethylation in sperm, and these are strongly enriched for annotations related to germ cell processes. Far fewer were selectively hypomethylated in ESCs, and these were not enriched in any particular annotation category. Promoters of genes retaining nucleosomes have recently been shown to be hypomethylated in human sperm (Hammoud et al., 2009), and both of these features have been proposed to aid rapid activation during development. We find that gene-associated hypomethylation in sperm can be extended to more than 70% of all annotated genes in both human and chimp. Among these we failed to find any enrichment for regulators of early development. Instead, it seems that promoter regions are generally identified and bookmarked in sperm (see Zaidi et al., 2010).

Distinct Processes of HMR Formation Shape Germ Cell and ESC Methylomes

Genome-wide, CpG sites seem to adopt a methylated state by default (Edwards et al., 2010). This raises the problem of precisely how regions that become HMRs are identified as such. Regions of hypomethylation at promoters have been correlated with regulatory DNA in various developmental contexts (Illingworth et al., 2008; Laurent et al., 2010; Rollins et al., 2006; Straussman et al., 2009). Based upon analysis of histone marks and on the proposed binding properties of DNMT3s (Dhayalan et al., 2010; Zhang et al., 2010), active transcription and accompanying methylation of K4 on histone H3 are thought to locally inhibit the methylation machinery. This could enable large-scale recognition of promoter regions if widespread transcription occurs during fetal germ cell development as genomic methylation patters are erased and reset. It is also plausible that specific protein/DNA complexes act locally even in the absence of active transcription, to prevent access by de novo methyltransferases. Proteins observed to function as boundary elements, such as CTCF and Sp1 (reviewed in Gaszner and Felsenfeld, 2006), provide candidates for such functions.

Despite overall similarity in the sets of promoters they mark, the HMRs observed at promoters in mature male germ cells usually extend beyond the boundaries of HMRs in ESCs when the two overlap. These wider HMRs do not seem to reflect less precision in HMR boundaries, as methylation differences across HMR boundaries are similar between sperm and ESCs. Because this "nested" HMR phenomenon is observed at so many promoters, it does not seem to be associated with the regulation of any specific genes during germ cell development. We have observed a clear increase in CpG content through the extended portion of these HMRs relative to the genome-wide average, suggesting that they have to some degree avoided pressure to decay and hence are more than a transient state. The phenomenon that we observe is similar to the concept of CpG shores (Doi et al., 2009). Perhaps the extended HMRs in germ cells presage the extent of "shores" that correlate with changes in gene expression.

Our data suggest that HMRs emerge from de novo methylation in male germ cells with sizes that differ from those that emerge from somatic reprogramming. Thus, despite involvement of similar methyltransferases and targeting of similar sets of sequences, the determinants of HMR sizes likely differ between the two reprogramming events. We have begun to see hints to the mechanisms determining such differences by comparing boundary-associated motifs in sperm and ESCs.

Transposon Hypomethylation in Sperm

It is thought that germ cell genomes must be closely guarded from the activity of mobile genetic elements. Although repeats were generally heavily methylated, we did find HMRs that overlapped repeats, and these were substantially more prevalent in sperm. We and others have characterized a conserved, small RNA-based silencing pathway, termed the piRNA pathway, that is important for recognizing and silencing mobile elements in germ cells (Aravin and Hannon, 2008). Our data indicate that both individual element copies and broader element subfamilies can evade piRNA-based silencing. Yet, both these element copies and element families are often efficiently silenced during preimplantation development. This suggests fundamental differences in the mechanisms that recognize repeats and mark them for repression during the two major waves of epigenetic reprogramming in mammals.

Examining patterns of repeat-associated HMRs is potentially enlightening. HMRs are more prevalent in younger transposon subfamilies, and the hypomethylated regions themselves tend to overlap with promoters or regulatory regions, just as they do in genes. Thus, it may be that active elements evade default methylation by being initially recognized as gene-like as a consequence of their binding transcription factors and possibly even being transcribed. In these cases, we imagine that silencing of most elements would be enforced by the piRNA pathway but that some sites, such as those we observe herein, might still escape. A number of examples can be cited in support of this hypothesis. The 5′ untranslated regions (UTRs) of the L1PA subfamilies are known to carry conserved YY1-binding sites, whereas other recent subfamilies acquired RUNX3- and SRY-binding motifs, all of which could promote transcription in developing germ cells (Khan et al., 2006; Lee et al., 2010). Similarly, the sperm-enriched hypomethylated EVR9 LTR12 elements have been shown to bind NF-Y, MZF1, and GATA-2 in erythroid K562 cells (Yu et al., 2005). In each of these cases, HMRs within these elements tend to encompass such potential transcription factor-binding sites.

Similarly, Alu RNAs have been detected in human sperm (Kochanek et al., 1993). This suggests a potential link between Alu HMRs and the transcriptional activity of individual repeats, though previous studies also reported that the binding of SABP across Alu elements in sperm prevents their methylation (Chesnokov and Schmid, 1995). Interestingly, Alu hypomethylation is not seen in female germ cells (Liu et al., 1994) and has been proposed as one mediator of sex-specific imprints.

Centromeric Satellite Methylation

Satellites resist methylation in sperm when localized in clusters at centromeres but are generally methylated when located elsewhere even if they are clustered. This is consistent with previous observations made in mouse through the use of methylation-sensitive enzymes (Yamagata et al., 2007). Recent reports have shown that the transient transcriptional activation of paternal pericentromeric satellites was essential for centromeric heterochromatin formation in two-cell zygotes (Probst et al., 2010). This could indicate that hypomethylation of satellite repeats in male germ cell marks paternal centromeres, in a manner similar to imprinting, allowing their rapid transcriptional activation upon fertilization.

In addition to a characteristic location within chromocenters in sperm, centromeres display a distinct chromatin structure differentiating them regionally during meiosis from other chromosomal regions (reviewed by Dalal, 2009). This has prompted suggestions that centromeric chromatin states might be critical for proper meiosis, a hypothesis strongly supported by our observation of selective hypomethylation of megabase domains of centromeric satellite clusters. Prior studies have demonstrated that derepression of satellite repeats in mitotic cells creates segregation defects due to the formation of anaphase bridges (Frescas et al., 2008). Low methylation levels have also been correlated with the ability to bind cohesin complexes (Parelho et al., 2008). Considered as a whole, these observations suggest a model in which selective hypomethylation of centromeric satellites might be critical for accurate chromosome segregation during meiosis.

Differential Repeat Methylation between Species

The most striking example of species-specific methylation to emerge from our analysis involved the SVA elements. These primate-specific composite elements contain a high density of CpGs, remain active in human and chimp, and include many copies that are clear orthologs between human and chimp (Bantysh and Buzdin, 2009; Mills et al., 2006). Transduction of SVAs has been implicated in human diseases and gene formation (Damert et al., 2009; Ostertag et al., 2003). Our results indicate that for a subset of SVA elements, the ability to methylate these elements has either been acquired along the chimp lineage or lost in the human lineage during the past 6 million years, despite very little sequence change in these elements.

Mutual Canalization of the Genome and the Epigenome

It has been thought that CGIs arose as the result of protection from methylation-associated deamination over long evolutionary periods. This is consistent with the observed correlation between the location of CGIs and regions that lack methylation in both germline and somatic cells. However, recent results have pointed to functions for CGIs that may be associated with their high CpG density (Thomson et al., 2010), with the plausible interpretation that selection may be acting to preserve CpG density in CGIs. We find that although most CGIs fall within HMRs of sperm, most HMRs extend well beyond the annotated CGIs, even using weaker CGI definitions. Thus, hypomethylated regions in male germ cells do not appear to require a critical CpG density to avoid methylation. Instead, our results are consistent with CGIs arising as a consequence of different mutational pressures rather than selection for CpG density.

In our datasets, signatures of deamination-induced CpG depletion are clear. Yet we also observe CpG depletion from many sperm and ESC HMRs. Several scenarios could resolve this conundrum. For example, such regions may have been methylated for substantial periods prior to assuming their unmethylated status. Thus, they may have decayed at some time in the past but are now stabilized by their hypomethylated status. Such sites could also actually be methylated during a period of germ cell development to which our current datasets are blind (e.g., in fetal gonocytes or female germ cells). In accord with this explanation, we have observed distinct CpG densities associated with sperm-specific and ESC-specific HMRs. Moreover, at HMRs where the only central, nested portion is hypomethylated in ESCs, we observe greater CpG retention through regions hypomethylated in both ESCs and sperm. Overall, we cannot exclude a model in which selection acts to preserve critical functions requiring specific local CpG densities. However, our results lend additional support to recent conclusions of Cohen et al. (2011), whose sophisticated evolutionary modeling showed that CGIs can be explained without invoking selection on CpG sites. Our results suggest a refinement of the hypo-deamination model in which CpG retention is a function of the time spent hypomethylated during each generation in germ cells and their somatic precursors.

The detailed comparative analysis performed here has revealed that, over the ~6 million years since the divergence of human and chimp, most patterns of DNA methylation remain conserved in male germ cells. We have directly related evolutionary changes in CpG methylation with loss of CpG dinucleotides and have shown that even small differences in methylation can lead to substantial loss of CpGs over relatively short evolutionary periods. At the same time, there are many genomic regions that are highly conserved in sequence yet show quite different patterns of methylation. This could indicate an ability of the genome and the epigenome to evolve independently. However, we do find that the most drastic changes in methylation between human and chimp, where an HMR in one species shows high levels of methylation in the other, are accompanied by an increased sequence divergence even at non-CpG dinucleotides. One interpretation is that most species-specific HMRs have arisen newly along one lineage with these novel functional elements showing signs of recent adaptation. On the other hand, if this accelerated sequence change were more a reflection of relaxed selective pressure, we would expect species-specific HMRs to more frequently result from loss of functional elements along the opposite lineage. Resolution of these questions can only come from a broadening to many more species of the studies reported herein.

EXPERIMENTAL PROCEDURES

Detailed methods can be found in the Extended Experimental Procedures.

Sperm Collection

Two anonymous human donors were used and data pooled after sequencing. Two chimp donors were used. Semen was collected at the New Iberia Research Center (New Liberia, LA) or the Southwest National Primate Research Center (San Antonio, TX, USA). Coagulated semen was separated from the liquid phase manually. Both human and chimp samples were diluted (1:1) in HBS buffer (0.01M HEPES, ph 7.4; 150 mM NaCl) and passed though a silica-based gradient, SpermFilter (Cryobiosystems), by centrifugation (according to manufacturer's instructions).

Library Preparation

DNA from ~100 million cells was extracted and sheared to a size of ~150–200 nt by sonication. Double-stranded DNA fragments were end repaired, A-tailed, and ligated to methylated Illumina adaptors. Ligated fragments were bisulfite converted using the EZ-DNA Methylation-Gold Kit (Zymo research). Following PCR enrichment, fragments of 340 to 360 bp were size selected and sequenced.

Computational Methods

Reads were mapped with RMAPBS (Smith et al., 2009). The accuracy of our mapping method is discussed in the Extended Experimental Procedures. Mapped reads were used to infer the methylation frequency at each CpG dinucleotide. These frequencies, along with the number of reads contributing to each frequency estimate, were supplied to a segmentation algorithm used to identify HMRs. Ortholog mapping between human and chimp was done with the liftOver tool available through the UCSC Genome Browser. Sequence conservation between human, chimp, and was measured based on MULTIZ 44-way vertebrate alignments, also available through the UCSC Genome Browser. Complete details of all computational methods are provided in the Extended Experimental Procedures.

ACCESSION NUMBERS

Data analyzed herein have been deposited in GEO with accession GSE30340.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and five tables and can be found with this article online at doi:10.1016/j.cell.2011.08.016.

ACKNOWLEDGMENTS

We thank Michelle Rooks, Pramod Thekkat, and Colin Malone for help with experimental procedures and Assaf Gordon, Luigi Manna, and the CSHL and USC High Performance Computing Centers for computational support. We thank Babette Fontenot (New Iberia Research Center) and Jerilyn Pecotte (Southwest National Primate Center) for help with chimp sperm collection. We thank Sergey Nuzhdin, Ed Green, Peter Calabrese, Maren Friesen, Magnus Norborg, and Marie-Stanislas Remigereau for helpful discussions. This work was supported in part by grants from the NIH (R01HG005238 and 1RC2HD064459) and by a kind gift from Kathryn W. Davis.

Received: December 16, 2010
Revised: May 9, 2011
Accepted: August 10, 2011
Published: September 15, 2011

REFERENCES

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al; International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Aravin, A.A., and Hannon, G.J. (2008). Small RNA silencing pathways in germ and stem cells. Cold Spring Harb. Symp. Quant. Biol. *73*, 283–290.

Bantysh, O.B., and Buzdin, A.A. (2009). Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. Biochemistry (Mosc.) *74*, 1393–1399.

Bestor, T.H. (1998). Cytosine methylation and the unequal developmental potentials of the oocyte and sperm genomes. Am. J. Hum. Genet. *62*, 1269–1273.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96–99.

Chesnokov, I.N., and Schmid, C.W. (1995). Specific Alu binding protein from human sperm chromatin prevents DNA methylation. J. Biol. Chem. *270*, 18539–18542.

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

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.

Cooper, D.N., and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. *83*, 181–188.

Curley, J.P., Mashoodh, R., and Champagne, F.A. (2011). Epigenetics and the origins of paternal effects. Horm. Behav. *59*, 306–314.

Dalal, Y. (2009). Epigenetic specification of centromeres. Biochem. Cell Biol. *87*, 273–282.

Damert, A., Raiz, J., Horn, A.V., Löwer, J., Wang, H., Xing, J., Batzer, M.A., Löwer, R., and Schumann, G.G. (2009). 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. *19*, 1992–2008.

Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R.Z., Ragozin, S., and Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J. Biol. Chem. *285*, 26114–26120.

Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. *41*, 1350–1353.

Duncan, B.K., and Miller, J.H. (1980). Mutagenic deamination of cytosine residues in DNA. Nature *287*, 560–561.

Edwards, J.R., O'Donnell, A.H., Rollins, R.A., Peckham, H.E., Lee, C., Milekic, M.H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., et al. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res. *20*, 972–980.

Ehrlich, M., Zhang, X.Y., and Inamdar, N.M. (1990). Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. Mutat. Res. *238*, 277–286.

Enard, W., Fassbender, A., Model, F., Adorján, P., Pääbo, S., and Olek, A. (2004). Differences in DNA methylation patterns between humans and chimpanzees. Curr. Biol. *14*, R148–R149.

Frescas, D., Guardavaccaro, D., Kuchay, S.M., Kato, H., Poleshko, A., Basrur, V., Elenitoba-Johnson, K.S., Katz, R.A., and Pagano, M. (2008). KDM2A represses transcription of centromeric satellite repeats and maintains the heterochromatic state. Cell Cycle *7*, 3539–3547.

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

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet. *7*, 703–713.

Goodier, J.L., and Kazazian, H.H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell *135*, 23–35.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.

Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P., Park, J., Butler, J., Rafii, S., McCombie, W.R., Smith, A.D., and Hannon, G.J. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol. Cell. Published online September 15 2011. 10.1016/j.cell.2008.06.028.

Illingworth, R., Kerr, A., Desousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., et al. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol. *6*, e22.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. *16*, 78–87.

Kochanek, S., Renz, D., and Doerfler, W. (1993). DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J. *12*, 1141–1151.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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

Lee, S.H., Cho, S.Y., Shannon, M.F., Fan, J., and Rangasamy, D. (2010). The impact of CpG island on defining transcriptional activation of the mouse L1 retrotransposable elements. PLoS ONE *5*, e11353.

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

Liu, W.M., Maraia, R.J., Rubin, C.M., and Schmid, C.W. (1994). Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Res. *22*, 1087–1095.

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

Mills, R.E., Bennett, E.A., Iskow, R.C., Luttig, C.T., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). Recently mobilized transposons in the human and chimpanzee genomes. Am. J. Hum. Genet. *78*, 671–679.

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

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

Ostertag, E.M., Goodier, J.L., Zhang, Y., and Kazazian, H.H., Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. *73*, 1444–1451.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422–433.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

Probst, A.V., Okamoto, I., Casanova, M., El Marjou, F., Le Baccon, P., and Almouzni, G. (2010). A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev. Cell *19*, 625–638.

Rollins, R.A., Haghighi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., and Bestor, T.H. (2006). Large-scale structure of genomic methylation patterns. Genome Res. *16*, 157–163.

Sasaki, H., and Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. Nat. Rev. Genet. *9*, 129–140.

Schmid, C.W. (1991). Human Alu subfamilies and their methylation revealed by blot hybridization. Nucleic Acids Res. *19*, 5613–5617.

Shen, L., Wu, L.C., Sanlioglu, S., Chen, R., Mendoza, A.R., Dangel, A.W., Carroll, M.C., Zipf, W.B., and Yu, C.Y. (1994). Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J. Biol. Chem. *269*, 8466–8476.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics *25*, 2841–2842.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. Nat. Struct. Mol. Biol. *16*, 564–571.

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

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat. Genet. *20*, 116–117.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA elements: a hominid-specific retroposon family. J. Mol. Biol. *354*, 994–1007.

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

Yamagata, K., Yamazaki, T., Miki, H., Ogonuki, N., Inoue, K., Ogura, A., and Baba, T. (2007). Centromeric DNA hypomethylation as an epigenetic signature discriminates between germ and somatic cell lineages. Dev. Biol. *312*, 419–426.

Yu, X., Zhu, X., Pi, W., Ling, J., Ko, L., Takeda, Y., and Tuan, D. (2005). The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. J. Biol. Chem. *280*, 35184–35194.

Zaidi, S.K., Young, D.W., Montecino, M.A., Lian, J.B., van Wijnen, A.J., Stein, J.L., and Stein, G.S. (2010). Mitotic bookmarking of genes: a novel dimension to epigenetic control. Nat. Rev. Genet. *11*, 583–589.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., Fischle, W., and Jeltsch, A. (2010). Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. Nucleic Acids Res. *38*, 4246–4253.

Supplemental Information

EXTENDED EXPERIMENTAL PROCEDURES

Mapping Reads

Reads were mapped using the RMAPBS program. Our pipeline first removed adaptor sequence from any reads, discarding any reads with fewer than 40 high-quality bases after the adaptor was removed (reads were required to have at least 10 bases of overlap with the adaptor for any part to be trimmed). Ends of paired-end reads were mapped separately, and because adaptors were ligated to fragments prior to bisulfite treatment, the first end of each paired-end read was mapped using T→C wild-cards, and the second end of each read was mapped allowing A→G wild-cards (for details, see Smith et al., 2009). We allowed up to 10 mismatches when mapping reads, though the average was substantially lower, and low-quality positions in reads were never counted as a mismatch (recall that at least 40 high-quality positions were required). For each read, the mapping location was determined to be the location with the fewest mismatches. Reads for which two locations had the minimum number of mismatches were considered to map ambiguously and discarded.

In sequencing from the same library preparation, when multiple reads mapped to the exact same location, which we refer to as duplicate reads, we assumed these represent the same original molecule (e.g., PCR products of the same fragment). We discarded all but one read in the case of duplicates and retained the one with the fewest mismatches. This step of removing duplicates was only done prior to combining data from different library preparations. For paired-end reads, after mapping ends separately, any pairs found to overlap (indicating the original fragment had length less than 202 bases) were collapsed to prevent counting the same information twice in later analysis.

The reference genomes used were the hg18 (human) and panTro2 (chimp) genomes downloaded from the UCSC Genome Browser, and we excluded alternate haplotype sequences and ‘‘random’’ sequences for human. For chimp we excluded ‘‘random’’ sequences and the ‘‘unassembled’’ chromosome.

Accuracy of the Mapping Method

We conducted a simulation experiment to determine the portion of reads expected to be mapped to incorrect locations using the mapping method described above. The simulation used parameters for the following values:

- *Number of reads*. We set this value to 1 M.
- *Read length*. We used a read length of 101 nt (corresponding to the majority of our sequencing runs).
- *Methylation level*. Each CpG in sampled reads was considered methylated with probability 0.7. Although this does not simulate a specific methylation level for any given genomic CpG, the effect on mapping accuracy is the same.
- *Bisulfite conversion*. We set the simulation bisulfite conversion rate to 0.98, meaning that 98% of Cs that were not simulated as methylated were converted to Ts.
- *Sequencing errors*. We set the maximum number of sequencing errors per reads to 10. Each simulated read had 10 positions for errors sampled at random (though not uniformly; see below) with replacement. Errors were introduced after simulated bisulfite conversion.
- *Error distribution*. We used the error probabilities produced by the sequencing instrument in a 101 nt sequencing run to calibrate the probabilities for simulated errors occurring at any given position in the read. This results in a greater proportion of errors at the 3′ ends of simulated reads.

The simulation was done with human genome assembly hg18 (from UCSC Genome Browser) excluding unassembled centromeric regions. Simulated reads were mapped back to the genome using the procedure described above. Of the 1 million reads, 939,605 mapped back uniquely (94%). The portion mapping back to their location of origin was 935582 (99.6%). Because of sampling error positions with replacement, along with the nonuniform distribution for error locations, the average number of mismatches was 4.6 per mapped read, substantially greater than the average number of mismatches in our data. From this we conclude that any error introduced into downstream analysis by reads mapped to incorrect locations is sufficiently small to be negligible.

Association between Sets of Genomic Regions and Annotations

We stratified measures about CpG content and methylation in genomic regions according to their association with certain genomic annotations as follows. First we defined these associations so that they partition the set of regions in question. In other words, our definitions ensured that no HMR would be associated with both a promoter and a repeat element, even though a repeat could clearly exist inside the promoter of a gene. Our definitions were as follows:

- Promoter: Any region that overlaps the interval within 1 Kb of the transcription start site (TSS).
- Gene-proximal: Any nonpromoter region that overlaps the interval starting 10 Kb upstream of a TSS or 10 Kb downstream of a transcription termination site.
- Intergenic repeat: Any nonpromoter, non-gene-proximal region that overlaps a repeat.
- Intergenic nonrepeat: Any nonpromoter, non-gene-proximal region that does not overlap a repeat.

Repeat Definitions

We analyzed the following classes of repeats: LINE, SINE, LTR, Satellite, DNA, RNA, SVA, tRNA, low-complexity, and simple repeats. This list includes most of the repeats annotated in the RepeatMasker track from the UCSC Genome Browser.

SVA Elements with Identifiable Orthologs

We used SVA annotations from UCSC Genome Browser, which are based on RepBase. These annotations are constructed by matching repeat consensus sequences to the reference genome (hg18 and panTro2). SVA elements were retained in human if:

(1) The interval covered by the human copy lifts over to chimp
(2) The lift over target (in chimp) lifts back to human
(3) The target when lifting back from chimp to human is the same as the original interval

The same criteria were applied to chimp. This set of SVA elements was used in Figure 4A. This highly conservative criteria allowed us to compare methylation levels through copies of SVAs that existed in both species. The total number of these SVA copies included 358 pairs of high-confidence orthologs. The trends observed for this small, high-confidence set of elements is also reflected in the full sets of elements for human and chimp.

Calculation of Basic Statistics

Discarding Low-Quality Reads

Reads were first checked for the presence of adaptor sequence, indicating that the sequenced fragment was too short and sequencing proceeded into the adaptor at the other end of the fragment. We required at least a 10 base match starting from the beginning of the adaptor, excluding Ns in reads and allowing up to 2 mismatches. When such an adaptor sequence was found in a read, the read was trimmed after the beginning position of the match by replacing all subsequent bases (in the 3′ direction) with an N, which would not induce a mismatch during alignment. Any reads for which the final non-N base was at position 40 or less was discarded. Finally, any read with fewer than 28 non-N bases through its entire length of the read was discarded.

Estimating CpG Methylation Levels

For CpG *i*, define m_i as the number of reads showing methylation over position *i*, counting both strands. Define u_i as the number of reads showing lack of methylation over CpG *i*. The methylation level is estimated as $m_i/(m_i + u_i)$, which is an estimate of the probability that CpG *i* is methylated in a molecule sampled randomly from the cell population. Because CpG methylation is symmetric, m_i and u_i include observations associated with the cytosines on both strands for the *i-th* CpG.

Depth of Coverage and Bisulfite Conversion

All our measures of coverage are in terms of CpGs. Depth of coverage (fold coverage) is also measured only at CpGs and counts only T or C nucleotides (A or G for the second end of each read). Both these numbers are reflective of numbers calculated using all assembled bases. Bisulfite conversion is measured as the sum of the number of non-CpG cytosines that are converted to Ts (as indicated by Ts in reads mapping over non-CpG cytosines in the genome), divided by the total number of non-CpG cytosines in uniquely mapped reads.

Identifying Hypomethylated Regions

We identified hypomethylated regions (HMRs) using a stochastic segmentation to partition the methylome into alternating regions of hypermethylation and hypomethylation, the latter appearing as valleys in visual depictions of methylation profiles. More specifically, our method is based on a Hidden Markov Model (HMM; Durbin et al., 1999).

Our HMM consists of two states (for high and low methylation). To model the observations made at each individual CpG we use the following distributions. For a sequence of *n* CpGs in a contiguous chromosomal region, let p_i denote the true probability that CpG *i* is methylated in a molecule chosen at random from the sequenced sample. We assume that $p_i \sim \mathrm{Beta}(\alpha, \beta)$. The BS-seq data provides the numbers m_i and u_i of methylated and unmethylated reads, respectively, from which we estimate $\widehat{p}_i = m_i/(m_i + u_i)$. In calculating likelihoods of observations from a particular state (i.e., the emission distribution), we use a Beta-Binomial distribution. That is, we assume $m_i \sim \mathrm{BetaBinom}(\alpha, \beta, m_i + u_i)$, and

$$\Pr(m_i|\alpha,\ \beta,\ m_i+u_i) = \binom{m_i+u_i}{m_i} B(m_i+\alpha,\ u_i+\beta)/B(\alpha,\ \beta),$$

where *B* denotes the beta function. Critically, using this distribution allows us to model methylation probabilities accounting for the amount of data at each CpG while keeping the variance independent of the mean.

To fit distribution parameters for numerical convenience we work directly with the estimates $\widehat{p}_i$. This is because of the time required for maximum-likelihood computations directly with the Beta-Binomial. Instead, we estimate the maximum-likelihood parameters as though they were for a Beta distribution, and therefore satisfy

$$\psi(\widehat{\alpha}) - \psi(\widehat{\alpha}+\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\log(\widehat{p}_i)$$

and

$$\psi(\widehat{\beta}) - \psi(\widehat{\alpha}+\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\log(1-\widehat{p}_i)$$

with

$$\psi(x) = \frac{d}{dx}\log \Gamma(x).$$

To compute $\widehat{\alpha}$ and $\widehat{\beta}$, we use an iterative procedure. The initial parameter values are calculated as

$$\widehat{\alpha}^{(0)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\log(\widehat{p}_i)\right)$$

and

$$\widehat{\beta}^{(0)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\log(1-\widehat{p}_i)\right).$$

This initialization corresponds roughly to the assumption of $\alpha + \beta = 1$, as $\psi(1) = 0$. At each iteration, these estimates are updated using the formulas

$$\widehat{\alpha}^{(k)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\log(\widehat{p}_i) + \psi\left(\widehat{\alpha}^{(k-1)} + \widehat{\beta}^{(k-1)}\right)\right)$$

and

$$\widehat{\beta}^{(k)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\log(1-\widehat{p}_i) + \psi\left(\widehat{\alpha}^{(k-1)} + \widehat{\beta}^{(k-1)}\right)\right).$$

The inverse of the digamma (ψ) function can be calculated very easily by noting that $\psi^{-1}(x) = e^x + \epsilon$, for $0 \le \epsilon \le 1$ for any relevant values of x. We use a bisection search around e^x to evaluate ψ^{-1} and apply the iterative procedure until convergence criteria are satisfied.

After training the HMM parameters, HMRs were identified by posterior decoding, and then each was scored according to the sum of all $(1-\widehat{p}_i)$ for each CpG *i* in the HMR. Because a single CpG with an very high number of reads and a very low methylation level can theoretically be identified as a single-CpG HMR under our model, we included a procedure to identify only significant HMRs based on their score. The CpGs were randomly permuted, and then the random permutation was decoded to obtain an empirical distribution of random HMR scores. We obtained p values from this random distribution, and then applied the method of Benjamini and Hochberg (1995) to identify a cutoff for a false discovery rate (FDR) of 0.05. Finally, we retained as HMRs only those regions having a score more extreme than the identified 0.05 FDR cutoff.

Measuring Sequence Divergence and CpG Decay

We measured nucleotide-level conservation between human (hg18), chimp (panTro2), and gorilla (gorGor1) by using the MULTIZ 44-way alignment available through the UCSC Genome Browser (Blanchette et al., 2004). This alignment is referenced on human. Alignments for genomic intervals were extracted by identifying the blocks containing the start and end points of the region in human. If one of the two end-points was not found in the alignment, the region was determined not to be alignable. Positions in the alignments that correspond to gaps were not counted. A sequence was called "under decay" if it lost more than 5% of its CpGs; we required the inferred ancestral sequence to have at least 20 CpGs in order to make this determination.

Analysis of Nucleosome Retention Data

Nucleosome retention data was taken from Hammoud et al. (2009). Data from different donors for histone ChIP-seq experiments were pooled and mapped to the hg18 assembly using RMAP. Domains of retained nucleosomes and the H3K4me3 and H3K27me3 modifications were inferred using the RSEG algorithm (Song and Smith, 2011). This method identified 118318, 105150, and 193158 enriched domains for H3K4me3, H3K27me3, and retained histones, respectively.

Gene Ontology Analysis

To measure Gene Ontology category enrichment we used the web interface to the DAVID tool (Huang et al., 2008). For sperm and ESC-specific hypomethylated promoters we required that the promoter (−1 kb to +1 kb) overlap an HMR in one cell type, have a methylation level at least 0.5 in the other cell type, and have a difference of at least 2-fold between the lower and higher. We used RefSeq promoters downloaded from the UCSC Table Browser. To eliminate redundancy in the sets of Gene Ontology categories identified as enriched we used the REVIGO software through the web interface (Ŝkunca et al., 2009).

Motif Enrichment Analysis

We used programs for the CREAD package to analyze the HMR sequences for identifying enriched TFBS motifs. We used both libraries of known motifs from both TRANSFAC (Matys et al., 2006) and JASPAR (Sandelin et al., 2004). We measured enrichment relative to a randomly selected set of 5000 promoters from among those that had low methylation levels in both sperm and ESCs. To eliminate bias due to different CpG content, CpG dinucleotides were inserted (or deleted) randomly in the background sequence set to bring the level of CpG up to that in the foreground. When randomly removing CpGs, they were mutated to TpG or CpA. The enrichment was measured using the Binomial p value option in the motifclass program of CREAD.

Enrichment of Sequence Patterns at HMR Boundaries

To measure enrichment of sequence patterns at boundaries of nested and extended HMRs, we used only those HMRs where a sperm HMR fully contained exactly one ESC HMR. We only considered hexameric patterns that had a CpG dinucleotide at the center and no other CpG dinucleotides in order to avoid bias introduced by the fact that CpG content will differ on either side of an HMR boundary (which we already know). We determined the expected number of occurrences of a sequence pattern by counting the number of genomic CpGs centered on that pattern, and dividing by the number of genomic CpGs.

Use of Individual Variation Data from HapMap

Individual variation data from HapMap 3 (including phases II and III) were downloaded from http://hapmap.ncbi.nlm.nih.gov. We used the CEU population, as this most closely matched the sperm donors, and the amount of data was almost as high as any of the other 10 populations. In identifying sites to use, we took only sites where the HapMap annotated ancestral allele was at the C of a CpG site (on either strand), and we also required that at least 5 reads mapped over that CpG in our bisulfite sequencing data. We used Chi-squared goodness-of-fit tests to determine that the frequency spectra differed between low and high methylation levels for each type of derived nucleotide (A, G, or T).

SUPPLEMENTAL REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B *57*, 289–300.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708–715.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. (1999). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge, UK: Cambridge University Press).

Huang, D., Sherman, B., and Lempicki, R. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. Nucl. Acids Res. *34* (*suppl_1*), D108–D110.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open access database for eukaryotic transcription factor binding profiles. Nucl. Acids Res. *32*, D91–D94.

Ŝkunca, N., Ŝmuc, T., and Supek, F. (2009). REViGO: Redundancy Elimination and Visualization of Gene Ontology Term Lists. In The 3rd Adriatic Meeting on Computational Solutions in the Life Sciences.

Song, Q., and Smith, A.D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics *27*, 870–871.

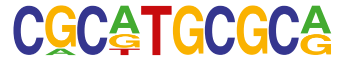
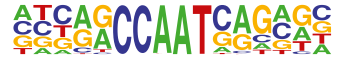
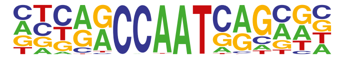
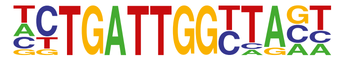
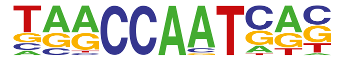
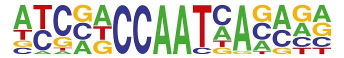
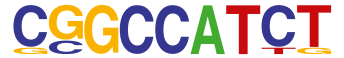
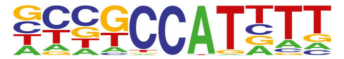
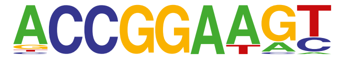
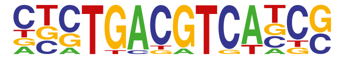
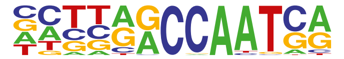
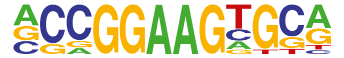
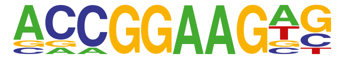
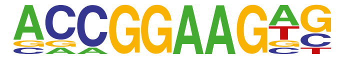
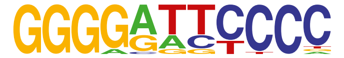
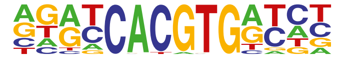
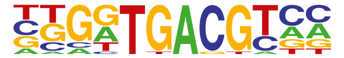
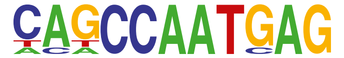
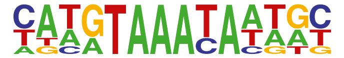
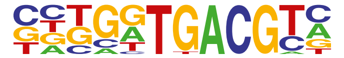
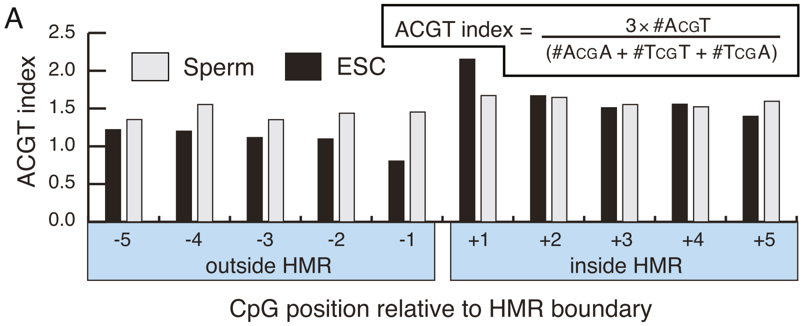
Factor	Logo	p-value
1. NRF1		5.34e-09
2. NFY/CP1/CBF/HAP2		6.39e-09
3. NFY/CP1/CBF/HAP2		8.45e-09
4. NFY/CP1/CBF/HAP2		6.83e-08
5. NFY/CP1/CBF/HAP2		5.16e-07
6. NFY/CP1/CBF/HAP2		8.18e-07
7. YY1/NF-μE1		1.59e-06
8. YY1/NF-μE1		2.46e-05
9. ETS		3.41e-05
10. CREB/ATF		1.13e-04
11. NFY/CP1/CBF/HAP2		1.83e-04
12. ETS		1.92e-04
13. ETS		2.16e-04
14. ETS		2.16e-04
15. NF-κB		3.29e-04
16. EBOX2		3.30e-04
17. CREB/ATF		3.55e-04
18. NFY/CP1/CBF/HAP2		3.59e-04
19. FOX		3.67e-04
20. CREB/ATF		4.04e-04

Figure S1. Related to Figure 2

Transcription factor-binding site motif enrichment associated HMRs overlapping promoters in human sperm but not in ESCs. p values of enriched motifs were calculated using a random subset of HMRs overlapping promoters in both cell types as a background.



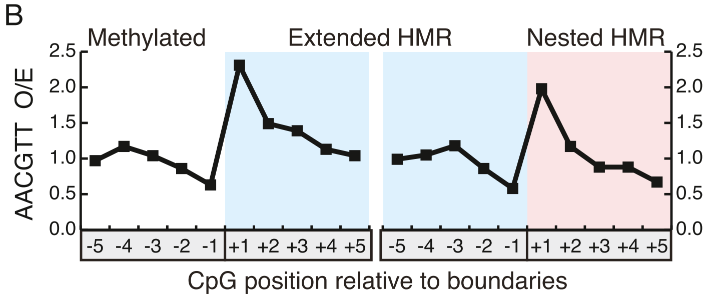


Figure S2. Related to Figure 3

(A) The AcgT index measured at CpG sites surrounding HMR boundaries in sperm (gray bars) and ESCs (black bars). Each data point corresponds to a CpG at positions −5 to +5 relative to HMRs boundaries.

(B) Observed-to-expected ratio for occurrences of the AACGTT pattern at each of the CpG positions from −5 to +5 relative to the boundaries of nested ESC and extended sperm HMRs.

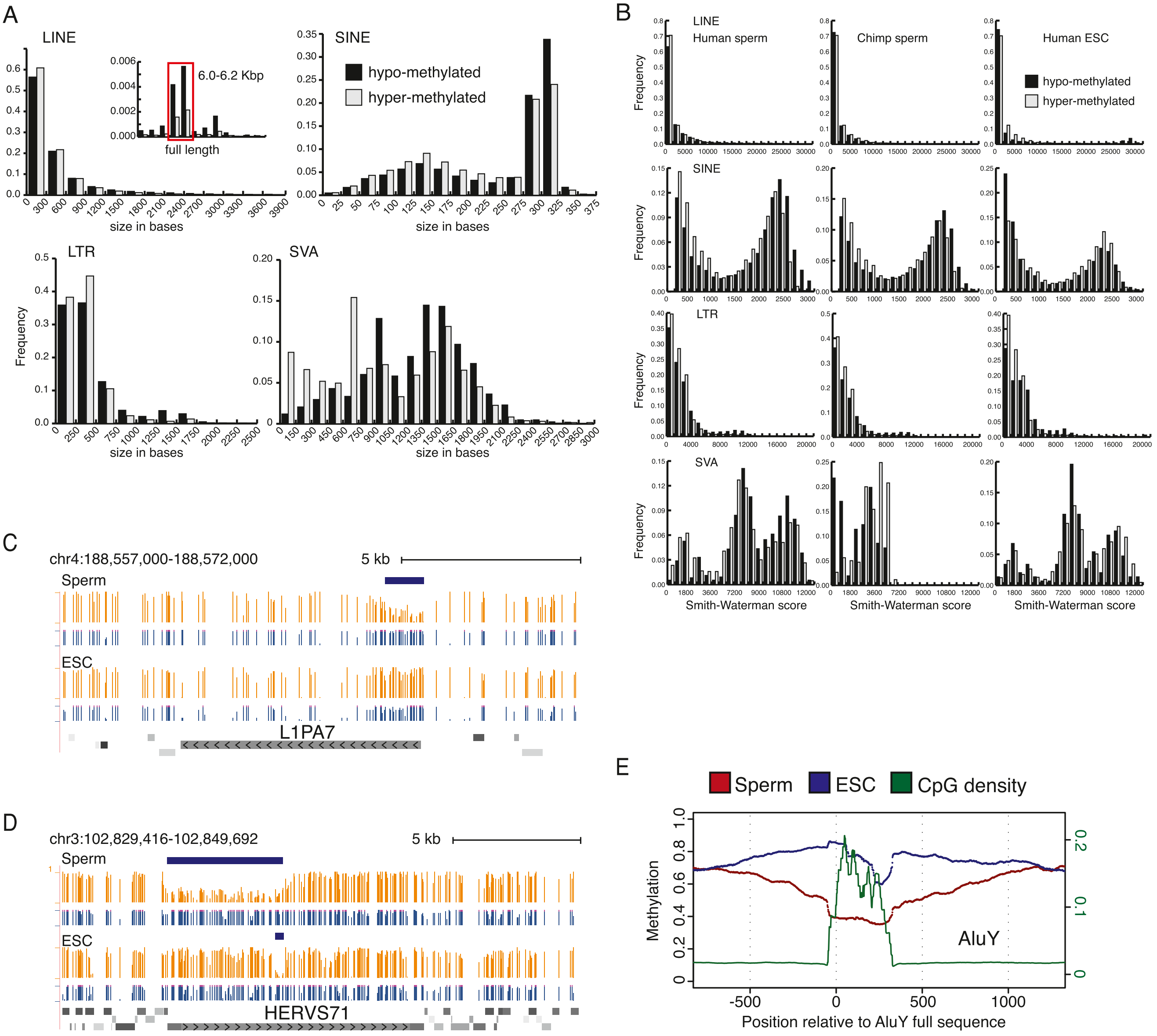


Figure S3. Related to Figure 4

(A) Size distribution of retrotransposons that are hypomethylated (black) and methylated (white) in human sperm. For each bin, the frequency of element copies is plotted.

(B) Histograms of Smith-Waterman scores of retro elements relative to their consensus sequences for hypomethylated and methylated copies. Separate histograms are given for LINE, SINE, LTR, and SVA elements, and for methylation status in human sperm, chimp sperm, and human ESCs.

(C) Browser tracks showing methylation (orange), read coverage (blue), and HMRs (blue bars) over a full-length LINE-1 element (L1PA7) hypomethylated in human sperm (upper tracks) but not in ESCs (bottom tracks).

(D) Browser track (as displayed in A) showing sperm-specific hypomethylation of the ERV HERVS71 in human sperm.

(E) Average methylation levels across all AluY SINE elements in human sperm (red) and ESCs (blue). CpG density is also shown in green. Methylation levels and CpG densities are also shown across flanking regions.

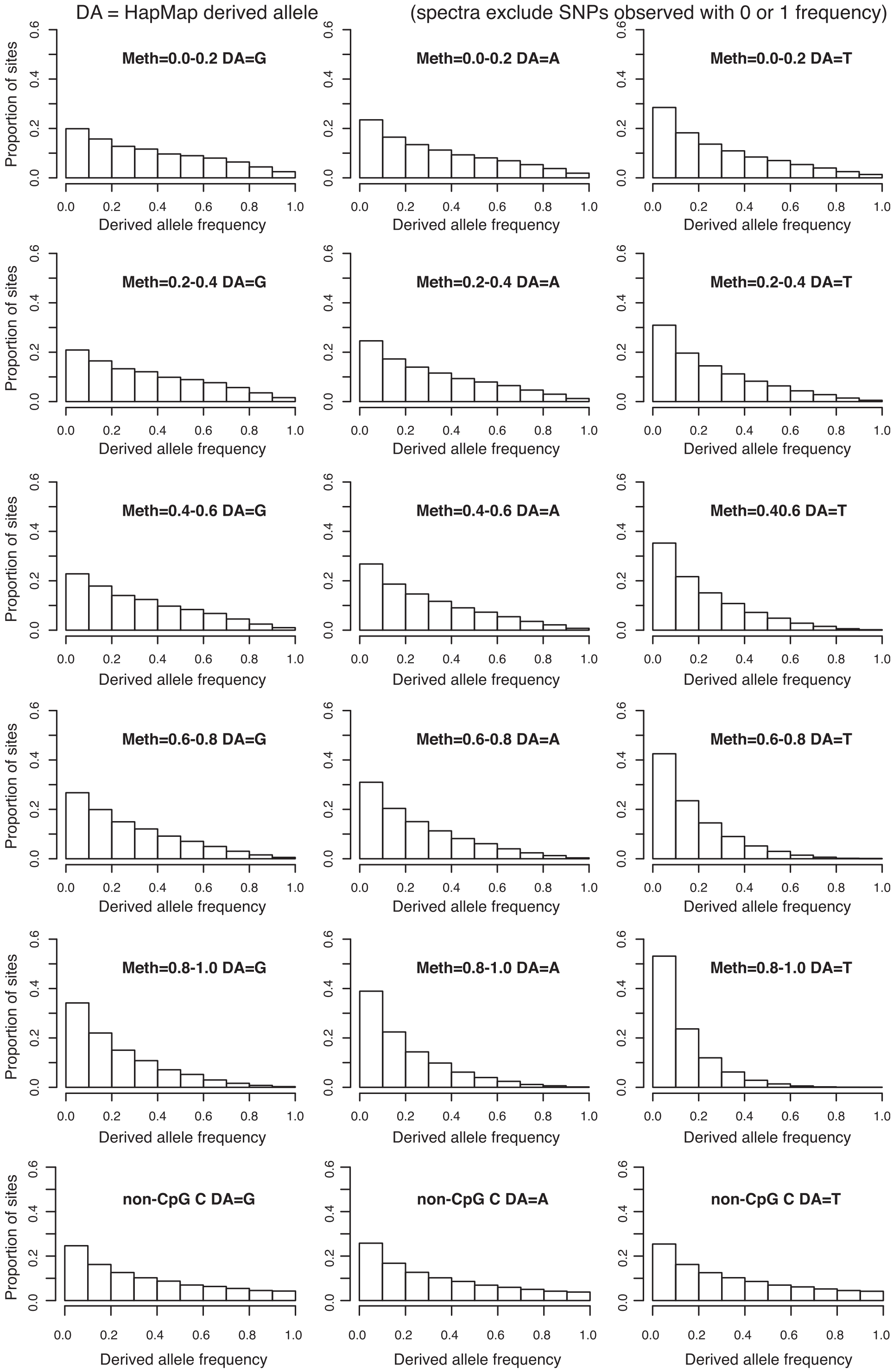


Figure S4. Related to Figure 6

Allele frequency spectra for each possible derived allele nucleotide at CpG sites treated symmetrically with cytosine as derived allele. For each derived allele, segregating sites were partitioned according to methylation levels in the intervals {[0.0, 0.2), [0.2, 0.4), … [0.8,1.0]}.