Supplementary information

Supplementary Figures Legends

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 (grey 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 ES cells. **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]}.

Supplementary Tables Legends

Tables S1. Related to Figure 1: A. Statistics on sizes, numbers and CpG contents of HMRs and CGIs. **B.** Association of human and chimp Sperm HMRs, as human ESC HMRs (Lister et al., 2009), with genome annotations. The distributions of HMRs are shown as a fraction of total HMRs and as a fraction of each annotation category. An individual HMR can overlap more than one annotation, and similarly a single annotation feature can overlap multiple HMRs.

Tables S2. Related to Figure 2: **A**. Gene Ontology analysis for genes with promoters having sperm-specific hypomethylation. GO term enrichment was calculated using the DAVID software (see supplementary methods for details). **B**. Non-redundant Gene Ontology terms for genes with promoters having sperm-specific hypomethylation. Non-redundant terms and dispensability indexes was measured using the REVIGO software (see supplementary methods for details). **C**. Gene Ontology analysis for genes with hypomethylated promoters in both human ESCs and Sperm. GO term enrichment was calculated using the DAVID software (see supplementary methods for details). **D**. Non-redundant Gene Ontology terms for genes with hypomethylated promoters in both human ESCs and Sperm. Non-redundant terms and dispensability indexes was measured using the REVIGO software (see supplementary methods for details).

Table S3. Related to Figure 3: Ratio of hexamer counts at +1 CpG (just inside HMR) over -1 CpG (just outside HMR), showing only the hexamers with the top 20 ratios for extended sperm HMRs and nested ESC HMRs.

Tables S4. Related to Figure 4: A. Satellite hypomethylation and methylation frequency according to centromeric status for human and chimp sperm, as well as human ES cells.

B. Coverage and mapping statistics over human and chimp retrotransposons, including LINE, LTR and SINE classes, as well as the SVA elements. Percent CpG covered refers to the percent of CpGs covered by at least one read. Methylation level by read is the average methylation from all reads mapping into elements from that family. Methylation by copy refers to the average of the mean methylation levels for each copy. The percent of copies with at least 5 informative reads refers to those copies for which at least 5 reads map over some CpG within the copy. **C.** Hypomethylation frequency

statistics for retrotransposon families in human sperm, chimp sperm and human ES cells. Columns labeled “copies” indicate the number of repeat copies for each family. For each of human sperm, human ES cells and chimp sperm, the total number of HMRs is indicated, as well as the number of hypomethylated copies, the expected number of hypomethylated copies, and the enrichment (hypomethylated copies divided by the expected). All annotations of repeats were taken from the Repeat Masker track in the UCSC Genome Browser. **D.** Hypomethylation frequency statistics for human LTR retrotransposon sub-families in human sperm and ES cells. Columns labeled “copies” indicate the number of repeat copies for each sub-family. The total number of HMRs is indicated, as well as the number of hypomethylated copies of the sub-family, the expected number of hypomethylated copies, and the enrichment (hypomethylated copies divided by the expected). All annotations of repeats were taken from the Repeat Masker track in the UCSC Genome Browser. **E.** Hypomethylation frequency statistics for human LINE retrotransposon sub-families in human sperm and ES cells. Columns labels and annotations as in E. **F.** Hypomethylation frequency statistics for human SINE retrotransposon sub-families in sperm and ES cells. Columns labels and annotations as in E. **G.** Hypomethylation frequency statistics for chimp LTR retrotransposon sub-families in chimp sperm. Columns labels and annotations as in E. **H.** Hypomethylation frequency statistics for chimp LINE retrotransposon sub-families in chimp sperm. **I.** Hypomethylation frequency statistics for chimp SINE retrotransposon sub-families in chimp sperm.

Tables S5. Related to Figure 6: **A**. Gene Ontology analysis for human sperm specific HMRs located within 10kb of a RefSeq promoter (no HMR in chimp sperm). GO term enrichment was calculated using the DAVID software (see supplementary methods for details).
B. Non-redundant Gene Ontology terms for human sperm specific HMRs located within 10kb of a RefSeq promoter (no HMR in chimp sperm). Non-redundant terms and dispensability indexes was measured using the REVIGO software (see supplementary methods for details).

Supplementary Experimental Procedures: Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates

Antoine Molaro[*,†] Emily Hodges[*,†] Fang Fang[‡] Qiang Song[‡]
W. Richard McCombie[*] Gregory J. Hannon[*,†,§] Andrew D. Smith [‡,§]

Contents

[*]Watson School of Biological Sciences
[†]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[‡]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[§]Corresponding author

1 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$\mapsto$C wild-cards, and the second end of each read was mapped allowing A$\mapsto$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 10M.
- *Read length.* We used a read length of 101nt (corresponding to the majority of our sequencing runs).
- *Methylation level.* Each CpG in sampled reads was considered methylated with probability 0.7. While 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 101nt 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 1M reads, 939605 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 non-uniform 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.

2 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 1Kb of the transcription start site (TSS).
- Gene-proximal: Any non-promoter region that overlaps the interval starting 10Kb upstream of a TSS or 10Kb downstream of a transcription termination site.
- Intergenic repeat: Any non-promoter, non-gene proximal region that overlaps a repeat.
- Intergenic non-repeat: Any non-promoter, non-gene proximal region that does not overlap a repeat.

3 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.

4 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 was applied to chimp. This set of SVA elements was used in Fig. 4(A). 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.

5 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.

6 Identifying hypomethylated regions (HMRs)

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 $p_i \sim \text{Beta}(\alpha, \beta)$. The BS-seq data provides the numbers m_i and u_i of methylated and unmethylated reads, respectively, from which we estimate $\hat{p}_i = m_i/(m_i+u_i)$. In calculating likelihoods of observations from a particular state (*i.e.* the emission distribution), we use a Beta-Binomial distribution. That is, we assume $m_i \sim \text{BetaBinom}(\alpha, \beta, m_i + u_i)$, 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 $\hat{p}_i$. 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

$$\psi(\hat{\alpha}) - \psi(\hat{\alpha}+\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\log(\hat{p}_i) \text{ and } \psi(\hat{\beta}) - \psi(\hat{\alpha}+\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\log(1-\hat{p}_i) \text{ with } \psi(x) = \frac{d}{dx}\log\Gamma(x).$$

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

$$\hat{\alpha}^{(0)} = \psi^{-1}\Big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log(\hat{p}_i)\Big) \quad \text{and} \quad \hat{\beta}^{(0)} = \psi^{-1}\Big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log(1-\hat{p}_i)\Big).$$

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

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

and

$$\hat{\beta}^{(k)} = \psi^{-1}\Big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log(1-\hat{p}_i) + \psi\big(\hat{\alpha}^{(k-1)} + \hat{\beta}^{(k-1)}\big)\Big).$$

The inverse of the digamma (ψ) function can be calculated very easily by noting that $\psi^{-1}(x) = e^x + \epsilon$, for $0 \leq \epsilon \leq 1$ 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 $(1 - \hat{p}_i)$ for each CpG i in the HMR. Since a single CpG with an very high number of reads and a very high 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 & 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.

7 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.

8 Analysis of nucleosome retention data

Nucleosome retention data was taken from Hammoud et al. (2009). Data from different donors for histone ChIP-seq experiments was 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 & Smith, 2011). This method identified 118318, 105150 and 193158 enriched domains for H3K4me3, H3K27me3 and retained histones, respectively.

9 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 ES cell-specific hypomethylated promoters we required that the promoter (-1kb to +1kb) 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 (Škunca et al., 2009).

10 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 (**?**) and JASPAR (**?**). We measured enrichment relative to a randomly selected set of 5000 promoters from among those that had low methylation levels in both sperm and ES cells. 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.

11 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.

12 Use of Individual Variation Data from HapMap

Individual variation data from HapMap 3 (including phase II and III) was 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).

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.* 14:708–715.

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

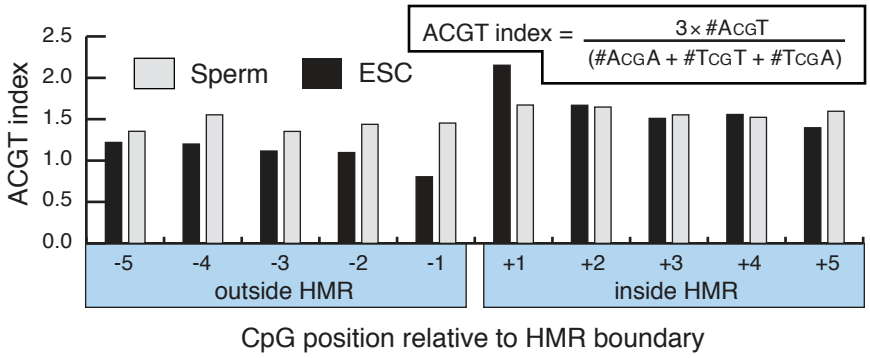
Huang D, Sherman B, Lempicki R (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4:44–57.

Škunca N, Šmuc T, 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*.

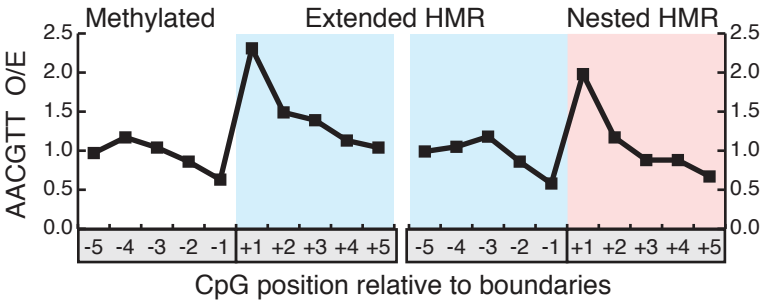
Song Q, Smith A (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* 27:870.

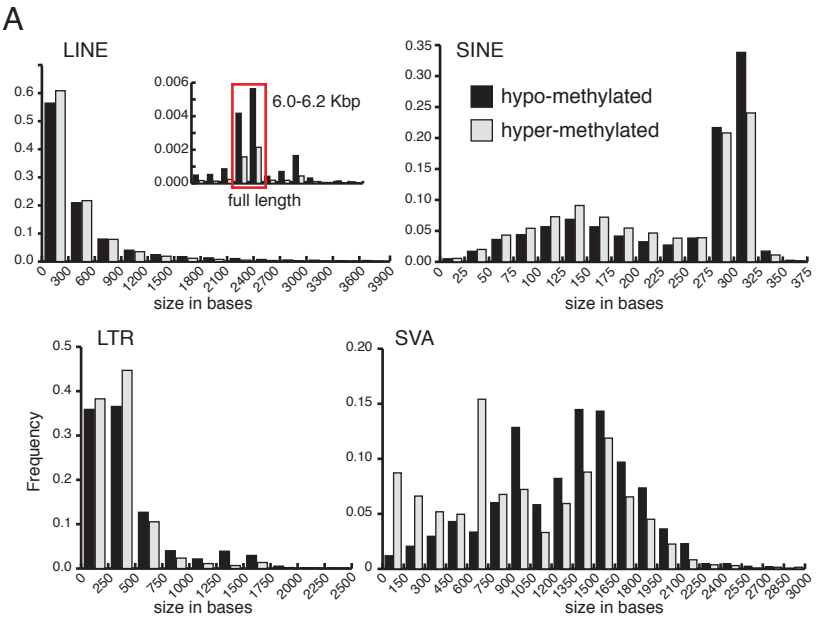
	Factor	Logo	p-value
1.	NRF1		5.34e-09
2.	NFY/CP1/CBF/HAP2		6.39e-09
3.	NFY/CP1/CBF/HAP2		8.45e-09
4.	NFY/CP1/CBF/HAP2		6.83e-08
5.	NFY/CP1/CBF/HAP2		5.16e-07
6.	NFY/CP1/CBF/HAP2		8.18e-07
7.	YY1/NF μE1		1.59e-06
8.	YY1/NF μE1		2.46e-05
9.	ETS		3.41e-05
10.	CREB/ATF		1.13e-04
11.	NFY/CP1/CBF/HAP2		1.83e-04
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

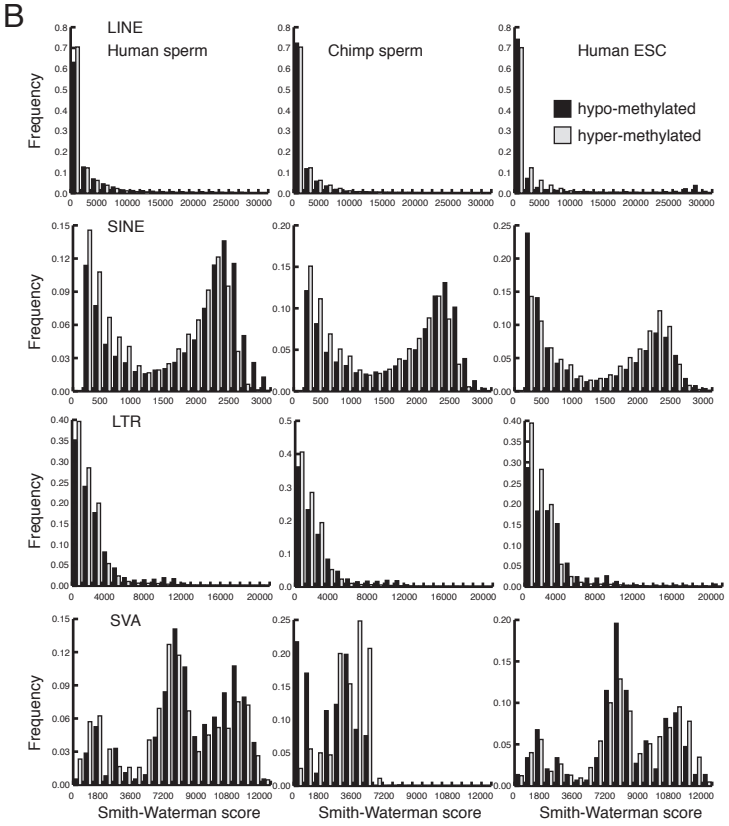
A

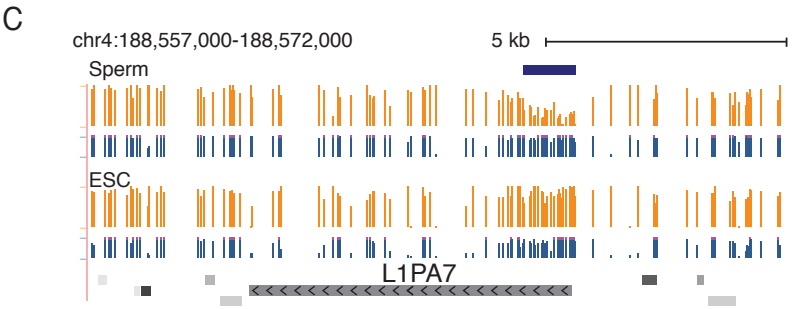

ACGT index = 3 × #ACGT / (#ACGA + #TCGT + #TCGA)
Sperm
ESC
ACGT index
2.5
2.0
1.5
1.0
0.5
0.0
-5 -4 -3 -2 -1
outside HMR
+1 +2 +3 +4 +5
inside HMR
CpG position relative to HMR boundary

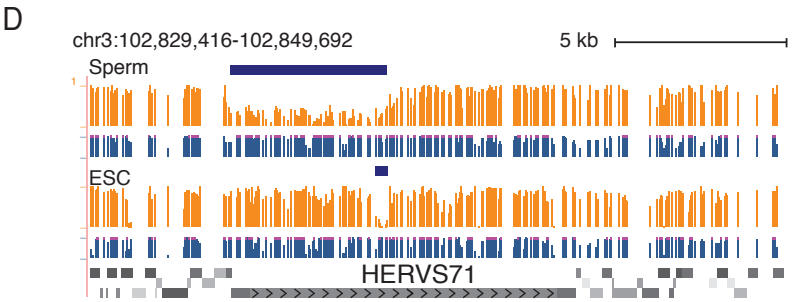
B

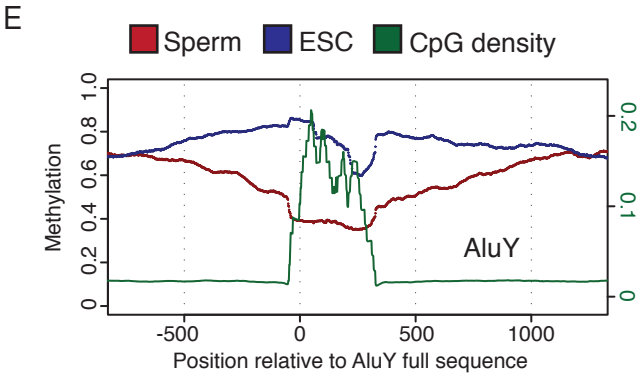

Methylated
Extended HMR
Nested HMR
AACGTT O/E
2.5
2.0
1.5
1.0
0.5
0.0
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
CpG position relative to boundaries

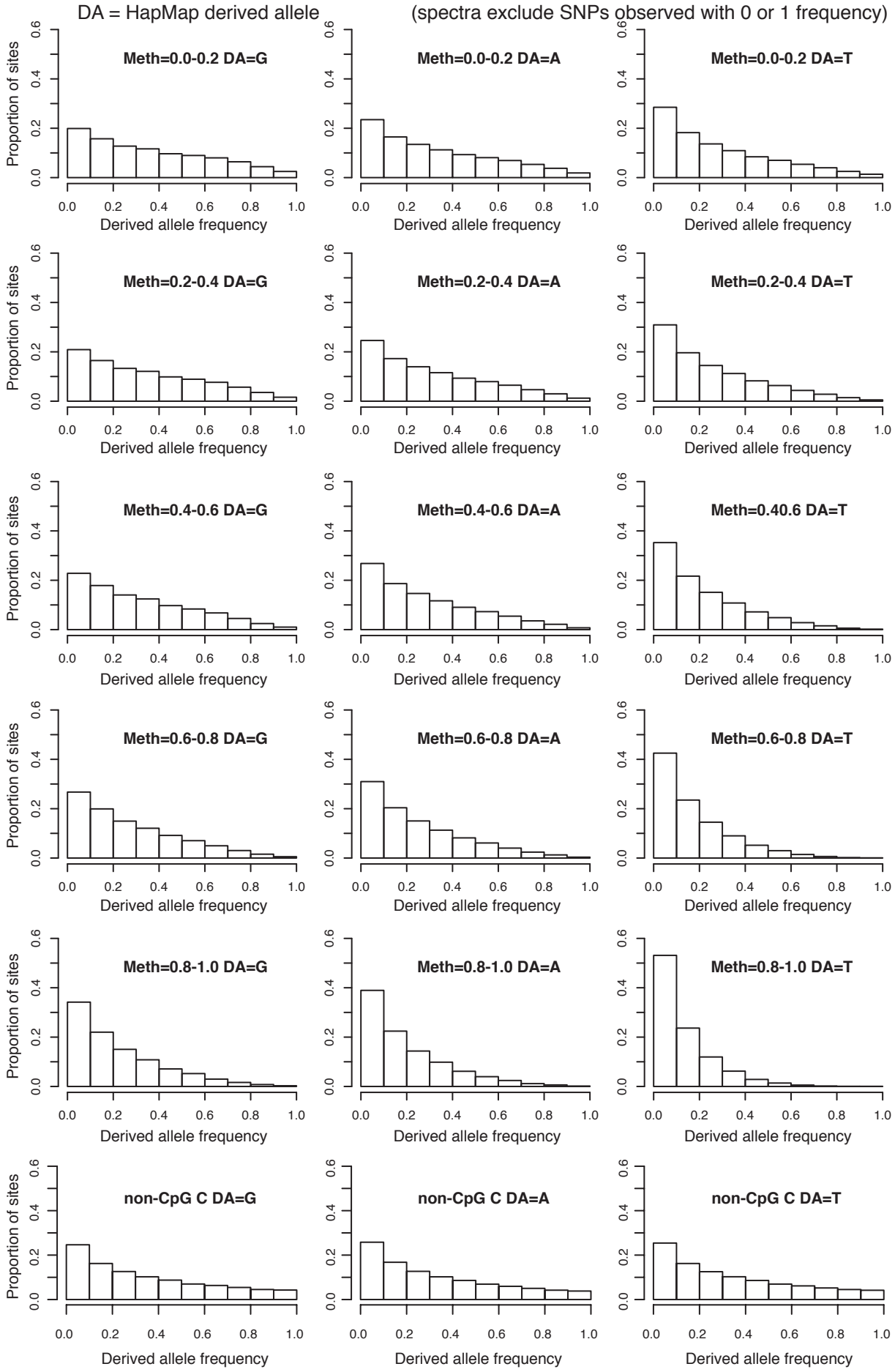

A
LINE
6.0-6.2 Kbp
full length
size in bases
SINE
hypo-methylated
hyper-methylated
size in bases
LTR
Frequency
size in bases
SVA
size in bases


B
LINE
Human sperm
Chimp sperm
Human ESC
hypo-methylated
hyper-methylated
Frequency
SINE
LTR
SVA
Smith-Waterman score
Smith-Waterman score
Smith-Waterman score


C
chr4:188,557,000-188,572,000
5 kb
Sperm
ESC
L1PA7


D
chr3:102,829,416-102,849,692
5 kb
Sperm
ESC
HERVS71


E
Sperm
ESC
CpG density
Methylation
AluY
Position relative to AluY full sequence


DA = HapMap derived allele
(spectra exclude SNPs observed with 0 or 1 frequency)
Meth=0.0-0.2 DA=G
Meth=0.0-0.2 DA=A
Meth=0.0-0.2 DA=T
Meth=0.2-0.4 DA=G
Meth=0.2-0.4 DA=A
Meth=0.2-0.4 DA=T
Meth=0.4-0.6 DA=G
Meth=0.4-0.6 DA=A
Meth=0.40.6 DA=T
Meth=0.6-0.8 DA=G
Meth=0.6-0.8 DA=A
Meth=0.6-0.8 DA=T
Meth=0.8-1.0 DA=G
Meth=0.8-1.0 DA=A
Meth=0.8-1.0 DA=T
non-CpG C DA=G
non-CpG C DA=A
non-CpG C DA=T
Proportion of sites
Derived allele frequency
0.0
0.2
0.4
0.6
0.8
1.0