Supplementary information: Adult tissue methylomes harbor epigenetic memory at embryonic enhancers

Gary C. Hon[1], Nisha Rajagopal[2], Yin Shen[1], David F. McCleary[1], Feng Yue[1], My D. Dang[1], and Bing Ren[1, 2, 3]

[1]Ludwig Institute for Cancer Research,

[2]Bioinformatics and Systems Biology Program,

[3]Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, and Moores Cancer Center, University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0653, USA

Supplementary Table Legends

Supplementary Table 1: tsDMRs in mouse tissues.

Supplementary Table 2: Motif analysis of tsDMRs not recovered by known regulatory elements.

Supplementary Table 3: AD-A and AD-I tsDMRs in mouse tissues.

Supplementary Table 4: Overlap of embryonic enhancers with adult AD-I tsDMRs.

Supplementary Table 5: Predicted enhancers.

Supplementary Table 6: Known motifs used in analysis.

Supplementary Note

Assessment of CpG density bias of tsDMRs

Identifying enhancers using tsDMRs requires the presence of CpG sequences, thus precluding the identification of enhancers with no CpGs (Supplementary Fig. 8c). To further clarify the relationship between CpG density and tsDMR detection, we observe that the enhancers and promoters with the greatest methylation variation are those with low CpG density (Supplementary Fig. 8a-b). Overall, we find that the statistical power of our method increases from 0.732 for tsDMRs with 3 CpGs to 0.997 for tsDMRs with at least 12 CpGs, with an average of 0.907 for all tsDMRs (Supplementary Fig. 8d).

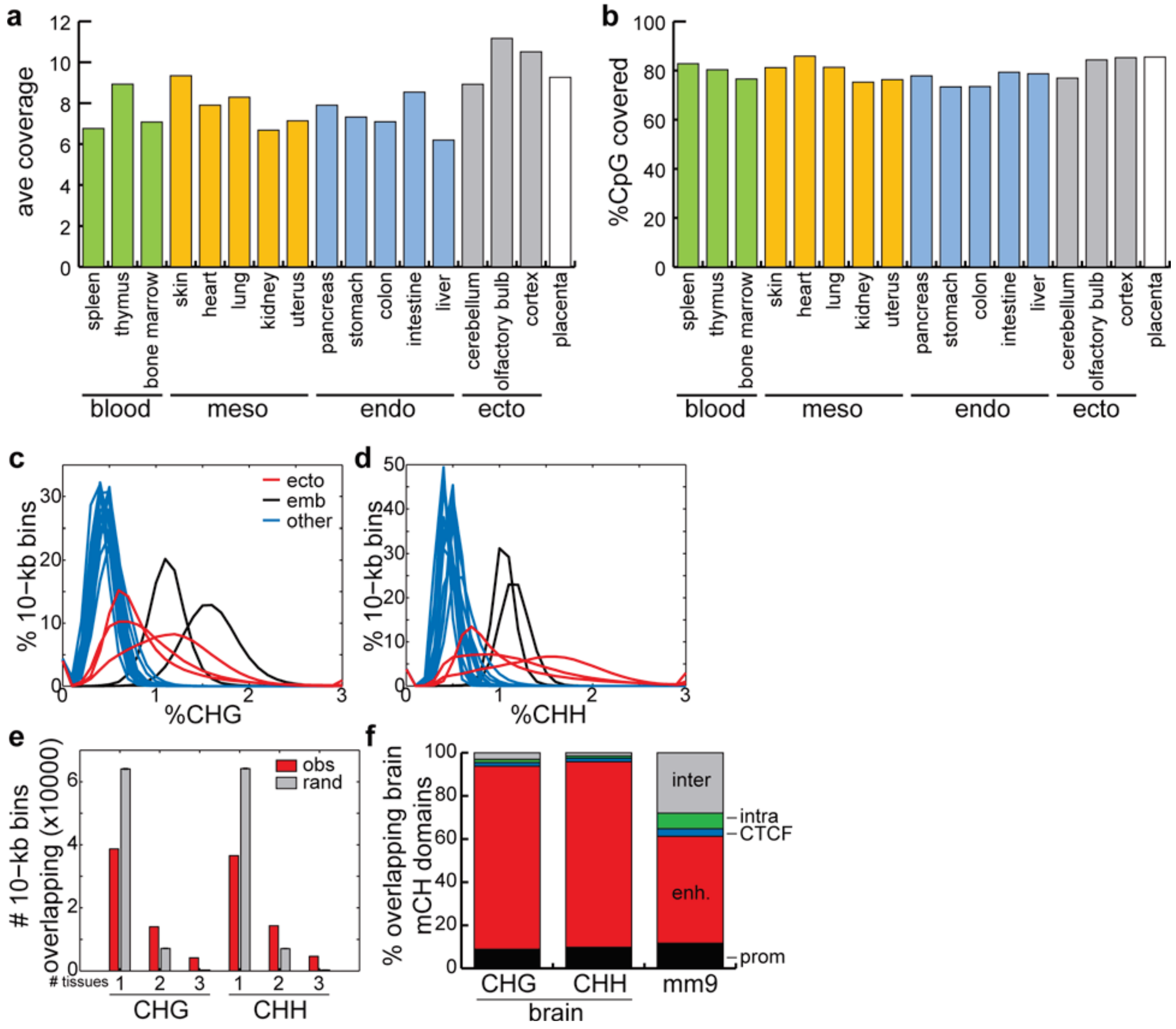
Due to limited sequencing depth, we are unable to identify differential methylation at the resolution of an individual cytosine. Rather, tsDMRs consist of groups of cytosines, which may bias toward higher density CpG regions. To address this point, we first examined the enrichment of promoters and enhancers at tsDMRs stratified by CpG content (Supplementary Fig. 8e-f). We find that while promoter enrichment increases with CpG content, enhancer enrichment exhibits a peak at 3.5% CpG content. The depletion at higher CpG content can be explained by the high CpG content of promoters. Regions of low CpG content are 19.0% depleted of enhancers compared to the peak enrichment. This depletion can be either explained by either an undersampling from our tsDMR approach or because enhancers are biased towards regions of intermediate CpG density.

To assess how differential methylation is behaving in other loci that are below the statistical power threshold of our sequencing depth and tsDMR approach, it is most appropriate to utilize high-depth bisulfite sequencing data. However, as this data currently does not exist in mouse cells, we turned to the human genome, for which our lab has recently published 5 DNA methylomes of ES-cell derived differentiation[1]. These maps are at a much higher sequencing depth than our study, with an average of 59X sequencing depth, and are thus considered more complete methylome maps. This higher depth allows us to identify base resolution cytosines with cell-type specific methylation. We computed the CpG density in 200-bp windows around these cytosines, and compared them with all cytosines in the human genome (Supplementary Fig. 8g, black curves). From this plot, we conclude that, if we had the sequencing depth to identify differential methylation at base resolution (dotted line black line), the CpG content of these bases is slightly greater than the genome-wide average (p < 1E-15, Wilcoxon). Next, turning to mouse, we compared the CpG content of tsDMRs identified in this study with the genome-wide expectation from the mouse genome (red curves). We find a more pronounced shift towards higher CpG content in tsDMRs (median = 2.5%) compared to the genome (median = 2.0%, p < 1E-15, Wilcoxon). Both of these results argue that sites of differential methylation, whether at base resolution or larger tsDMRs, display greater CpG content than the genome-wide expectation. However, tsDMRs display slightly more deviation from the genome than base resolution cytosines.

Next, we more precisely quantify this bias. By comparing the CpG content of tsDMRs with base resolution cytosines, we estimate a lower bound undersampling of low CpG content regions (<1.5% CpG) of 11%. Similarly, an upper bound of sampling bias due to limited sequencing depth can also be determined by comparing the CpG content of tsDMRs with the mouse genome-wide average, which we calculate to be 34%. Importantly, even if we assume that none of these low CpG density sites missed from our sampling are at enhancers, the fraction of tsDMRs at enhancers would only drop from 74% to 55%. Thus, even in the worst case scenario, we still estimate that the majority of differentially methylated cytosines are at enhancers.

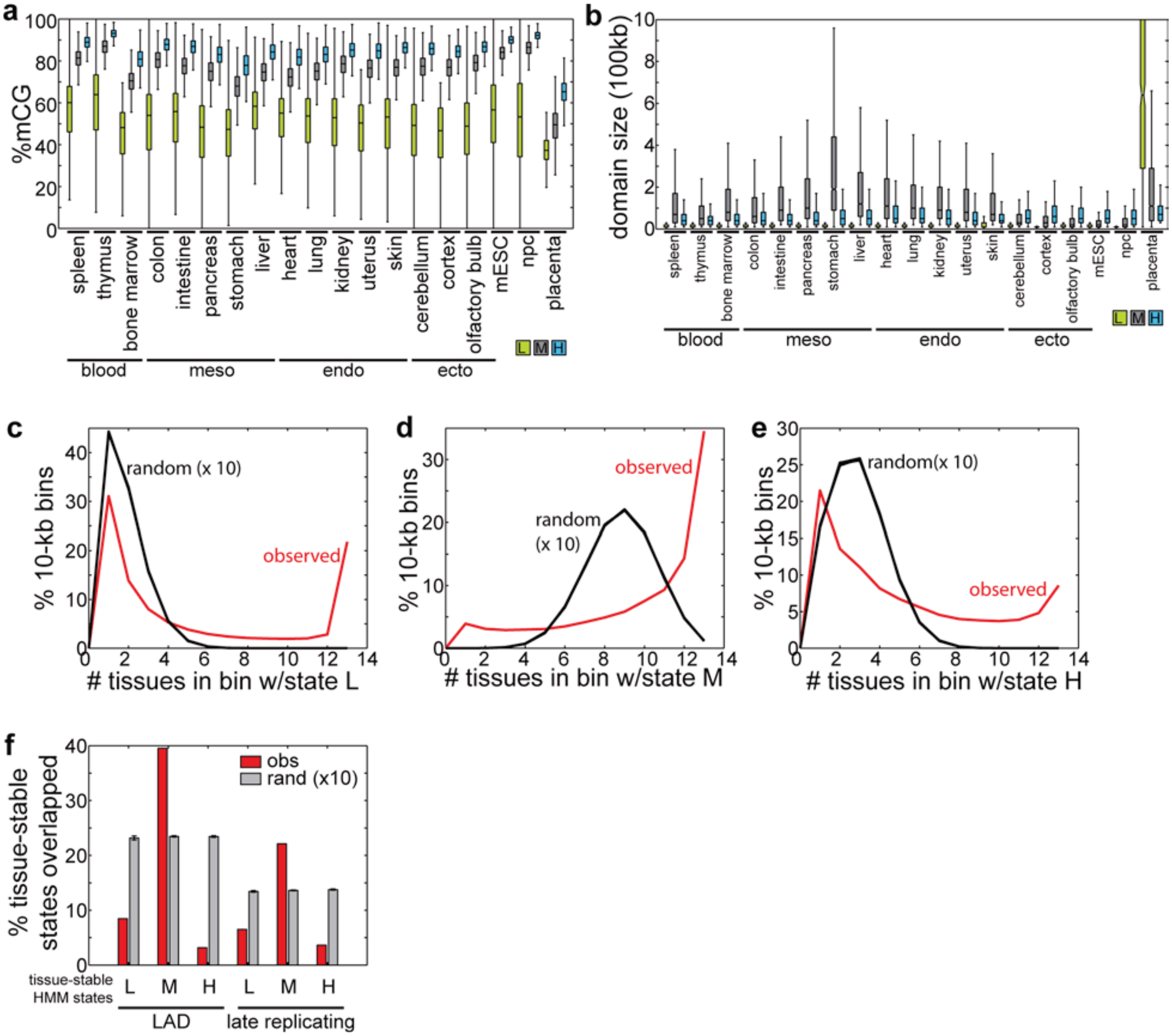
Future work using higher depth will be required to identify tsDMRs at CpG-poor regions and to assess their genomic function.

Supplementary Figures



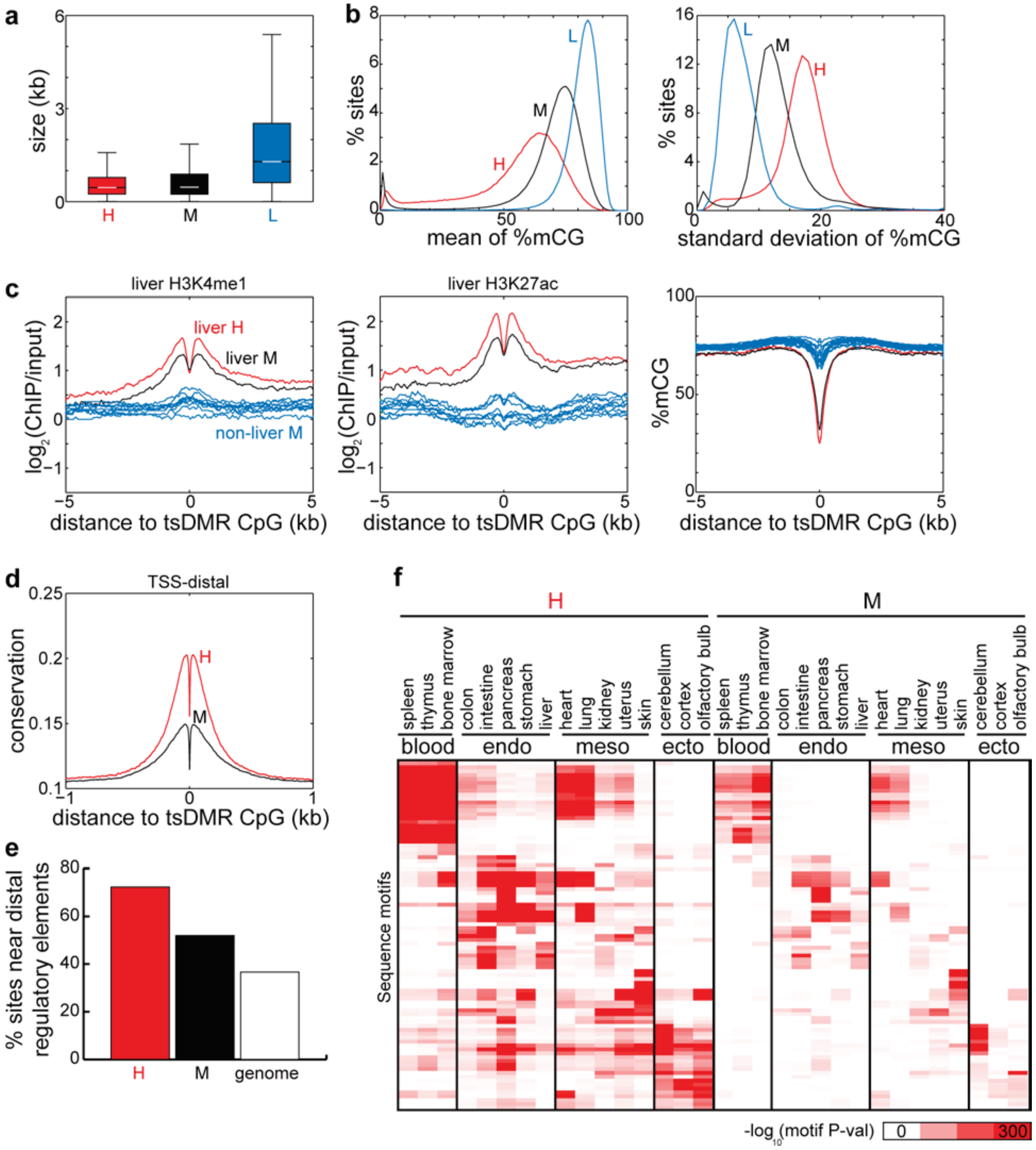
Supplementary Figure 1: Sequencing coverage of methylomes and analysis of non-CG methylation in mouse tissues

(**a**) Sequencing coverage of methylomes, relative to the size of the haploid mouse genome. (**b**) The percentage of CpG dinucleotides covered by bisulfite sequencing reads. (**c**) The global distribution of methylation in CHG context in brain tissues (red), embryonic cell lines (black), and other tissues (blue). H represents non-G nucleotides. (**d**) The global distribution of methylation in CHH context in brain tissues (red), embryonic cell lines (black), and other tissues (blue). H represents non-G nucleotides. (**e**) 10-kb domains enriched for non-CG methylation were overlapped between the three brain tissues. Shown are the number domains overlapped in exactly 1, 2, or 3 tissues (red). As a control, 10 sets of domains were randomly selected from the genome and overlapped (grey). Error bars indicate standard deviation. (**f**) The percentage of non-CG enriched domains overlapping with promoters, enhancers, CTCF binding sites, intragenic regions, and intergenic regions. The genome-wide average is indicated on the right.



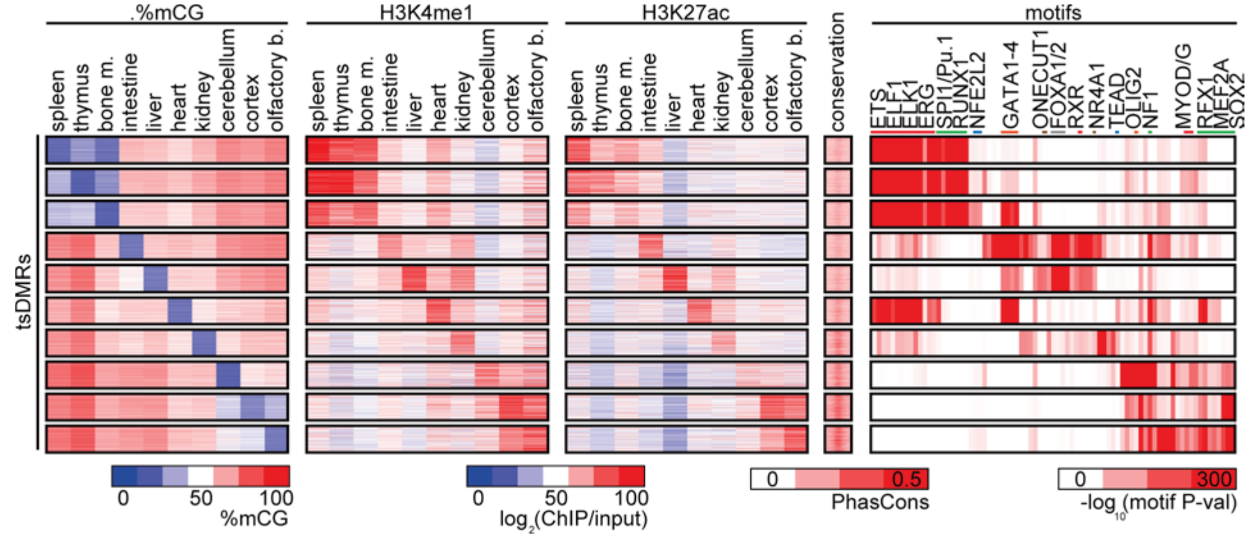
Supplementary Figure 2: Regions of intermediate DNA methylation are PMD precursors.

(**a**) The global distribution of CpG methylation for domains of low (L), medium (M), and high (H) methylation, as identified by HMMs trained on the genome segmented into 10-kb bins. A separate HMM was trained for each tissue. (**b**) The distribution of domain size for domains identified in (**a**). Domains of (**c**) low, (**d**) medium, or (**e**) high methylation were overlapped between all non-ectodermal tissue. Shown is the distribution of the number of tissues overlapped by domains (red). As a control, 10 sets of domains were randomly selected from the genome and overlapped (black). (**f**) Domains methylated at the same level (L/M/H) in all non-ectodermal tissues were overlapped with domains consistently lamina-associated (LAD, identified as the overlap of mouse ES cells, NPCs, astrocytes, and MEFs domains) and consistently late replicating (identified as the overlap of 22 mouse developmental cell lines with enrichment greater than -0.5). As a control, 10 sets of domains were randomly selected from the genome and overlapped (grey). Error bars indicate standard deviation. Boxplot edges indicate the 25th and 75th percentiles, and whiskers indicate non-outlier extremes.



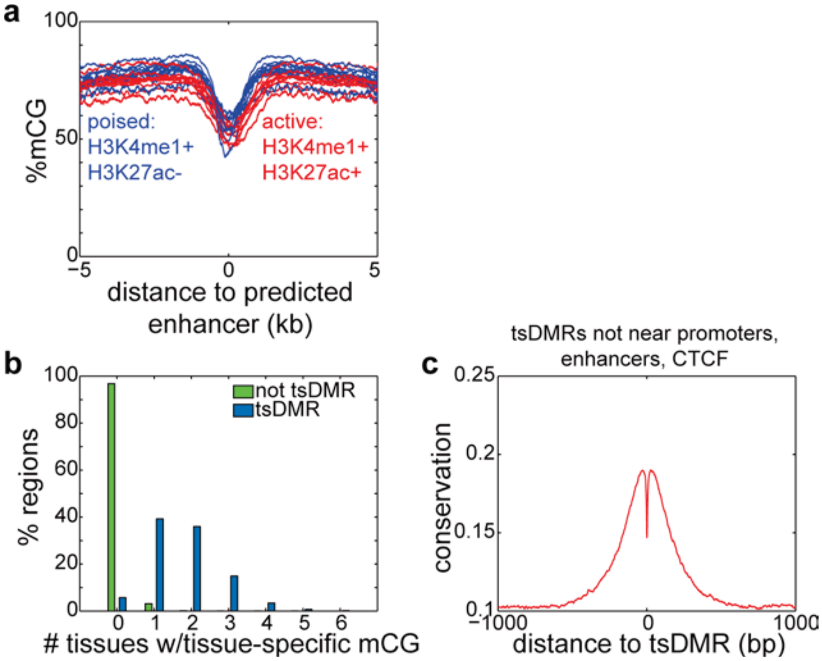
Supplementary Figure 3: Analysis of sites of intermediate methylation variance.

(**a**) The size distribution for regions of high (H, red), medium (M, black), and low (L, blue) tissue-specific methylation variance. Boxplot edges indicate the 25th and 75th percentiles, and whiskers indicate non-outlier extremes. (**b**) The distribution of average (left) and standard deviation (right) of %mCG, for H/M/L regions. (**c**) Enrichment of H3K4me1 (left), H3K27ac (middle), and DNA methylation (right) in liver cells for liver tsDMRs of high (red) and medium (black) tissue-specific variance, compared to tsDMRs identified in other tissues (blue). (**d**) Average PhastCons conservation scores relative for H/M tsDMRs distal (beyond 2.5-kb) to annotated transcription start sites (TSS). Higher values indicate more conservation. (**e**) The percentage of H/M tsDMRs within 500-bp to known distal regulatory elements. (**f**) Heatmap representing the enrichment of transcription factor binding motifs for the H/M tsDMRs identified in each tissue. Each row represents a motif (Supplementary Table 6).



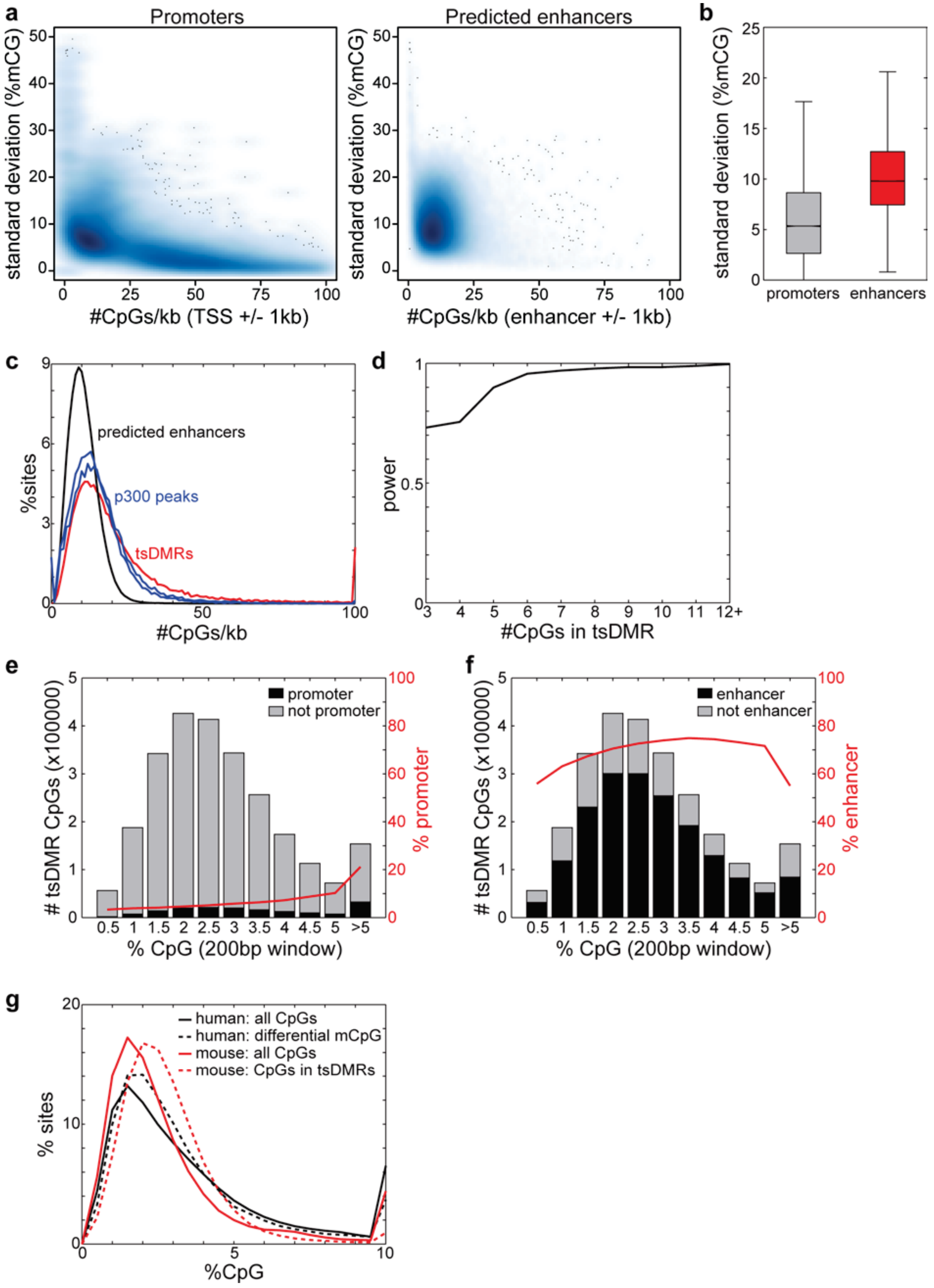
Supplementary Figure 4: Integrative analysis of methylation, histones, conservation, and motifs.

For tissues having both DNA methylation and histone modification data, shown (left to right) are heatmaps representing of enrichment of the following features at tsDMRs: DNA methylation, H3K4me1, H3K27ac, PhastCons evolutionary sequence conservation, and motifs.



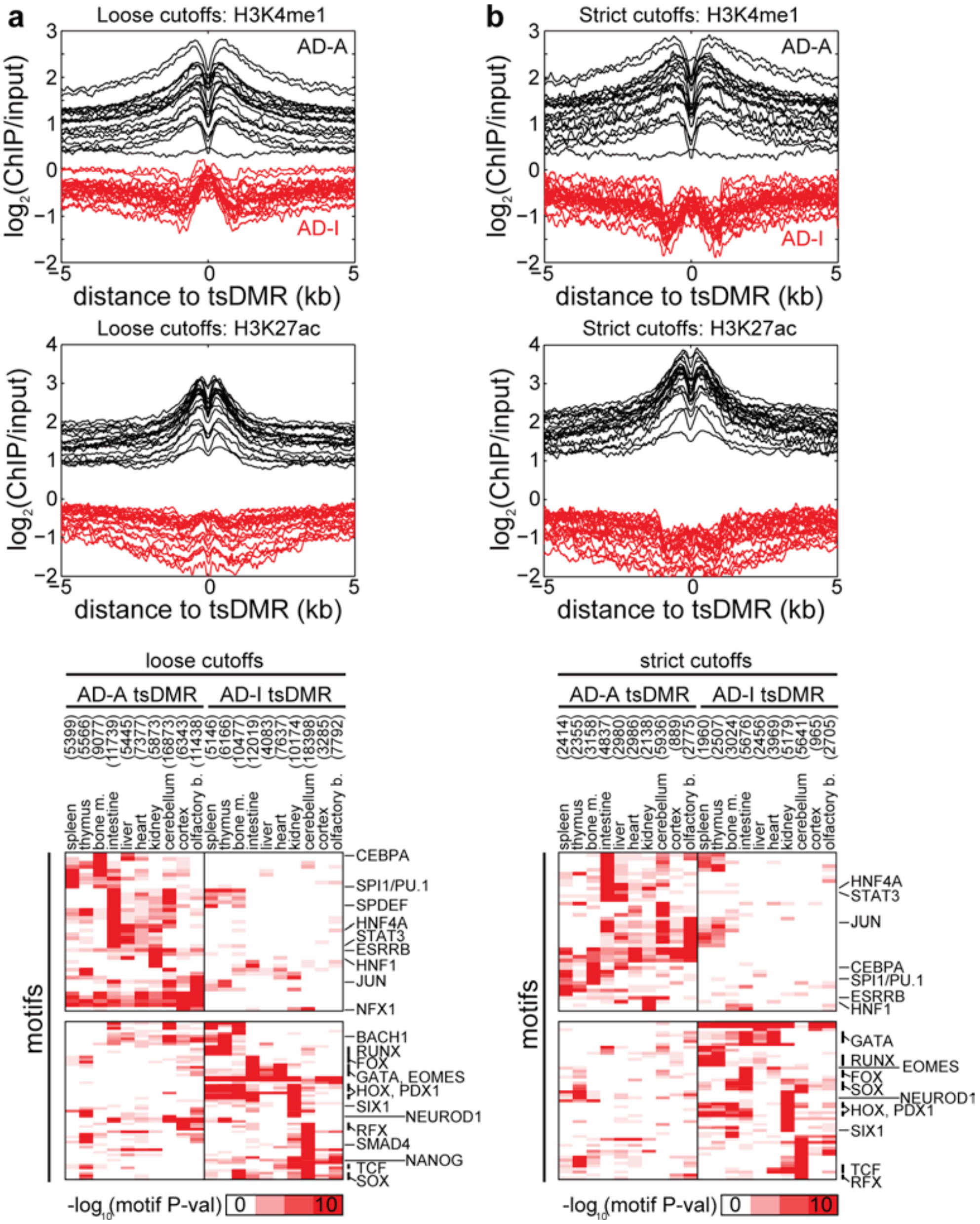
Supplementary Figure 5: Hypomethylation at poised enhancers, and overlap of tissue-specifically methylated regions.

(**a**) Profiles of DNA methylation at chromatin-predicted enhancers bearing an active (red) or poised (blue) chromatin state. (**b**) tsDMRs were associated with tissues with Shannon entropy, and shown is the distribution of number of tissues overlapped by tsDMRs (blue) and non tsDMRs (green). (**c**) Average PhastCons conservation scores of tsDMRs not recovered by promoters, enhancers, or CTCF sites. Higher values indicate more conservation.


a
Promoters
Predicted enhancers
standard deviation (%mCG)
0
10
20
30
40
50
#CpGs/kb (TSS +/- 1kb)
#CpGs/kb (enhancer +/- 1kb)
0
25
50
75
100
b
standard deviation (%mCG)
0
5
10
15
20
25
promoters
enhancers
c
%sites
0
3
6
9
predicted enhancers
p300 peaks
tsDMRs
#CpGs/kb
0
50
100
d
power
0
0.5
1
#CpGs in tsDMR
3
4
5
6
7
8
9
10
11
12+
e
# tsDMR CpGs (x100000)
0
1
2
3
4
5
promoter
not promoter
% promoter
0
20
40
60
80
100
% CpG (200bp window)
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
>5
f
# tsDMR CpGs (x100000)
enhancer
not enhancer
% enhancer
% CpG (200bp window)
g
% sites
0
10
20
human: all CpGs
human: differential mCpG
mouse: all CpGs
mouse: CpGs in tsDMRs
%CpG
0
5
10

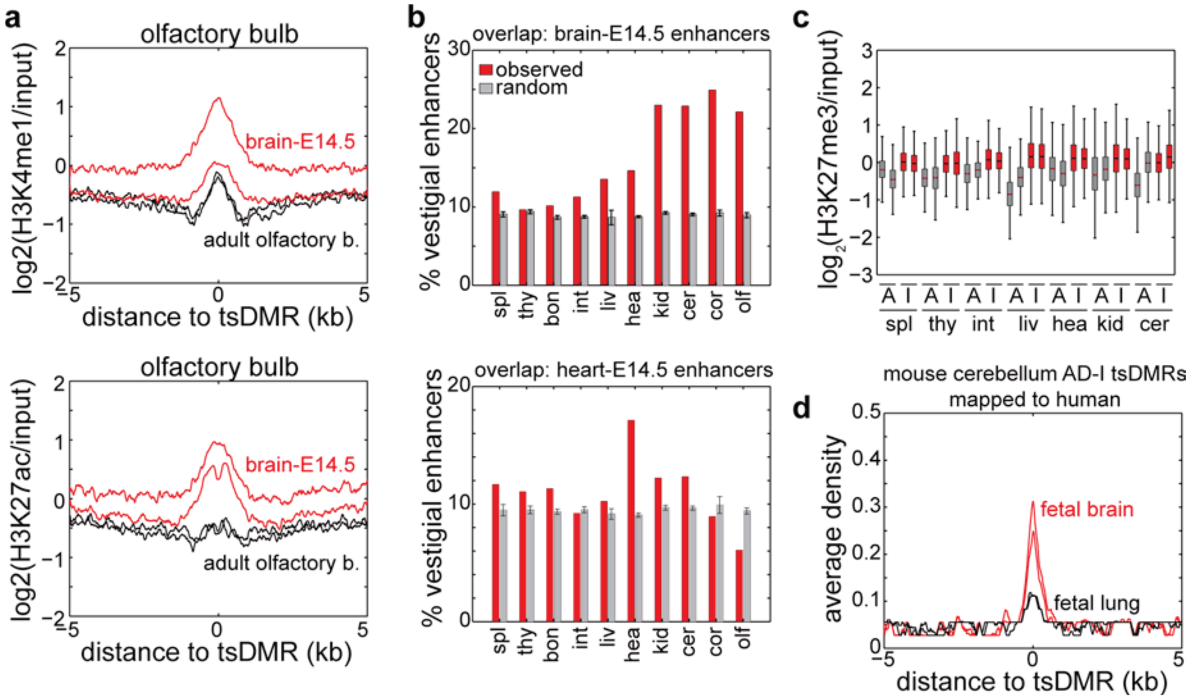
Supplementary Figure 6: CpG density analysis.

(**a**) Density plots illustrating the relationship between CpG density and tissue-specific methylation variation for promoters (known TSS +/- 1kb) and enhancers predicted from chromatin modifications (+/- 1kb). (**b**) Boxplots indicating the overall amount of tissue-specific variation for promoters and predicted enhancers. Boxplot edges indicate the 25th and 75th percentiles, and whiskers indicate non-outlier extremes. (**c**) The distribution of CpG density for predicted enhancers (black), two biological replicates of p300 binding sites in heart (blue), and tsDMRs (red). (**d**) Power as a function of the number of CpGs in tsDMRs. Power is estimated by comparing the distribution of Chi-Square scores in tsDMRs with the Chi-Square distribution (alpha = 0.05). (**e**) The CpGs within tsDMRs overlapping (black) or not overlapping (grey) promoters were stratified by CpG content. The bar chart indicates the number of CpGs in each CpG-density strata, and the percentage of promoter overlap is indicated in red. (**f**) As in (**e**), but for tsDMRs overlapping predicted enhancers. (**g**) Distributions of CpG content in 200-bp windows around all CpGs (solid lines) in human (black) and mouse (red). To identify base resolution cytosines that are cell-type specifically methylated in human, we used the χ^2 statistic described in this manuscript for individual CpGs on high-depth bisulfite sequencing experiments from human embryonic stem cells and four ES-derived cell lines[1]. Shown in dotted black is the subset of human CpGs exhibiting base resolution differential methylation. Shown in dotted red is the subset of mouse CpGs within tsDMRs identified in this manuscript.



Supplementary Figure 7: Robustness of vestigial enhancer analysis.

(**a**) Enrichment of H3K4me1 (top), H3K27ac (middle), and motifs (bottom) for vestigial enhancers defined with looser cutoffs than in the main text. For a given tissue, we defined AD-A tsDMRs as TSS-distal tsDMRs having $\log_2$(H3K27ac RPKM / input RPKM) ≥ 0.807 for either of two biological replicates, corresponding to a 1.75-fold ChIP enrichment over input. AD-I tsDMRs are defined as TSS-distal tsDMRs having $\log_2$(H3K4me1 RPKM / input RPKM) ≤ 0.41 and $\log_2$(H3K27ac RPKM / input RPKM) ≤ 0.41 for both biological replicates, corresponding to a maximum of 33% ChIP enrichment over input. (**b**) As in (**a**), but with stricter cutoffs. For a given tissue, we defined AD-A tsDMRs as TSS-distal tsDMRs having $\log_2$(H3K27ac RPKM / input RPKM) ≥ 2 for either of two biological replicates, corresponding to a 4-fold ChIP enrichment over input. AD-I tsDMRs are defined as TSS-distal tsDMRs having $\log_2$(H3K4me1 RPKM / input RPKM) ≤ 0 and $\log_2$(H3K27ac RPKM / input RPKM) ≤ 0 for both biological replicates, corresponding to a maximum of 0% ChIP enrichment over input.



Supplementary Figure 8: AD-I tsDMRs are active during development.

(**a**) For AD-I tsDMRs identified in adult olfactory bulb, shown is the average enrichment of H3K4me1 (top) and H3K27ac (bottom) in adult tissue (black) and embryonic day 14.5 tissue (red). Two biological replicates for each sample are shown. (**b**) Quantification of enrichment observed in Figure 7c for AD-I tsDMRs identified in adult tissues with enhancers predicted in E14.5 mouse brain and heart. As a comparison, overlap was also performed against random sets of AD-I tsDMRs (grey). Error bars indicate standard deviation. (**c**) Comparison of H3K27me3 enrichment in AD-A tsDMRs and AD-I tsDMRs in adult tissue. Spl, spleen; thy: thymus, int: intestine; liv: liver; hea: heart; kid: kidney; cer, cerebellum. Boxplots edges indicate the 25th and 75th percentiles, and whiskers indicate non-outlier extremes. (**d**) Mouse cerebellum AD-I tsDMRs were mapped to human (build hg19) and shown are average profiles of DNase I hypersensitivity in human fetal brain (red) compared to human fetal lung (black), each of which are represented by two biological replicates. These data were previously published and obtained from the Roadmap Epigenomics Project.

References

1. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-48 (2013).