Supplementary Information for
"Genome-wide maps of chromatin state in pluripotent and lineage-committed cells"

Tarjei S. Mikkelsen[1,2], Manching Ku[1,3], David B. Jaffe[1], Biju Issac[1,3], Erez Lieberman[1,2], Georgia Giannoukos[1], Pablo Alvarez[1], William Brockman[1], Tae-Kyung Kim[4], Richard P. Koche[1,2,3], William Lee[1], Eric Mendenhall[1,3], Aisling O'Donovan[3], Aviva Presser[1], Carsten Russ[1], Xiaohui Xie[1], Alexander Meissner[5], Marius Wernig[5], Rudolf Jaenisch[5], Chad Nusbaum[1], Eric S. Lander[1,5,*] and Bradley E. Bernstein[1,3,6,*]

1 Broad Institute of Harvard and MIT, Cambridge, MA 02139 USA

2 Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA

3 Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129 USA

4 Department of Neurology, Children's Hospital, Boston, MA 02115 USA

5 Whitehead Institute for Biomedical Research, MIT, Cambridge, MA 02139 USA

6 Department of Pathology, Harvard Medical School, Boston, MA 02115 USA

* These authors co-supervised the work.

Contents:

1 Supplementary Note
10 Supplementary Figures

Supplementary Tables and Data are available separately from the online version of the paper at www.nature.com/nature and from www.broad.mit.edu/seq_platform/chip/

Supplementary Note - ChIP-Seq read requirement, genome coverage and accuracy
The number of sequence reads required to map a chromatin feature can be estimated from a simple model.

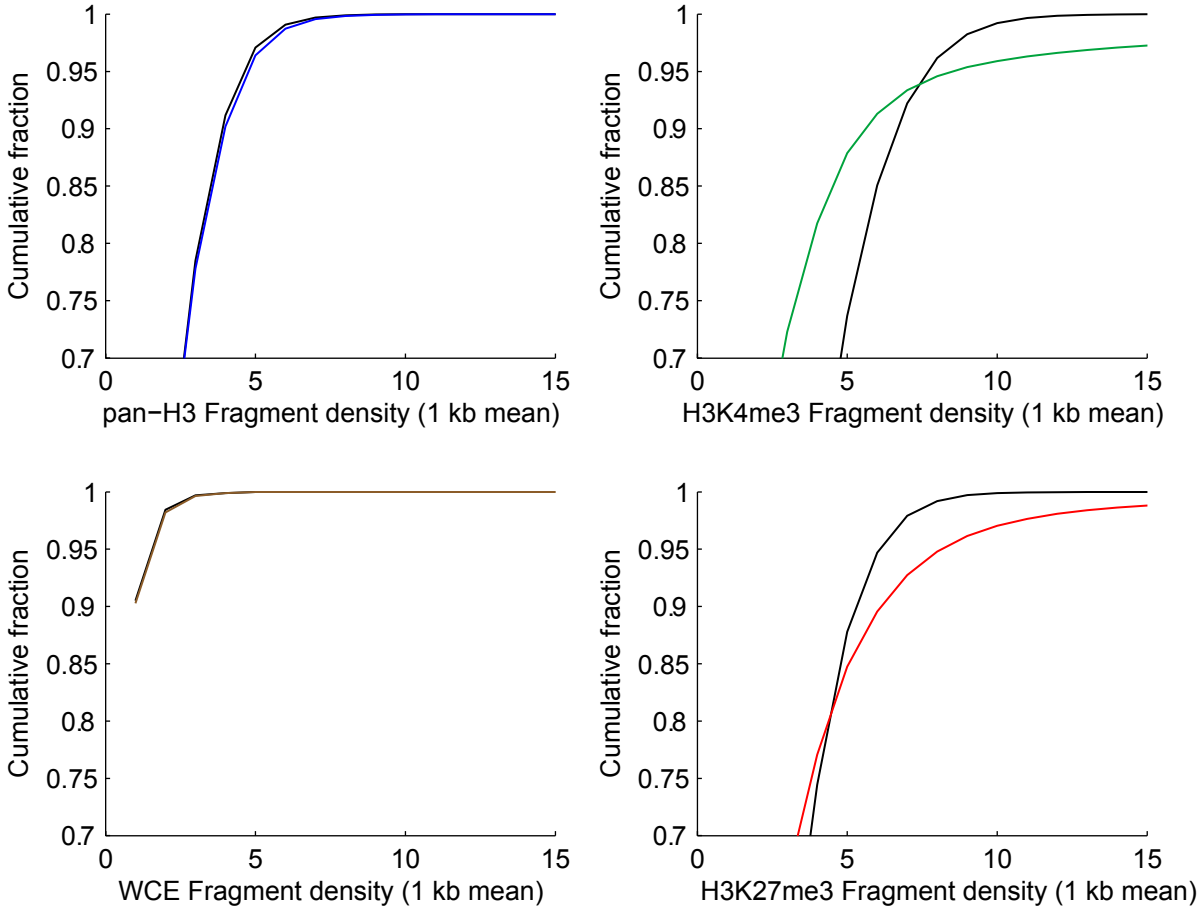
Suppose that the genome is divided into N non-overlapping bins of fixed size, that a fraction f of these bins contain a particular chromatin feature and that one performs ChIP-Seq with an antibody that enriches the sequence in these bins by a factor of e. If one collects a total of R sequence reads, the number of reads in a bin should approximately follow a Poisson distribution with mean eM for bins containing the feature and M for the other bins, where $M = R/N(ef+(1-f))$.

Theoretical specificity and sensitivity of ChIP-Seq, conditional on the number of reads, can be estimated from the overlap of the two distributions. For example, suppose that an epitope is present across 1% of the genome, and can be enriched 20-fold by an antibody. Mapping this epitope with 95% specificity and 95% sensitivity into bins of 500 bp would require ~2 million reads. Increasing the resolution to 200 base pairs would require ~5 million reads. Epitopes that enrich less efficiently require more reads (e.g. 10-fold enrichment and 200 base pair resolution would require ~10 million reads).

How much of the mouse genome can be interrogated by ChIP-Seq SMS reads? The proportion depends on the read length k and the mismatch tolerance d (where optimal read alignments are kept for analysis if they have no alternative alignment with ≤ d additional mismatches). In this report, we used k=27 and d=2 (although the actual read lengths varied from 27-36 bp). If we consider 500-bp windows in which at least half of the 27-mers are unique, then ~70% of all windows can be interrogated. Notably, this includes ~20% of all nucleotides in annotated interspersed repeats. Longer read lengths can provide over 80% theoretical coverage.

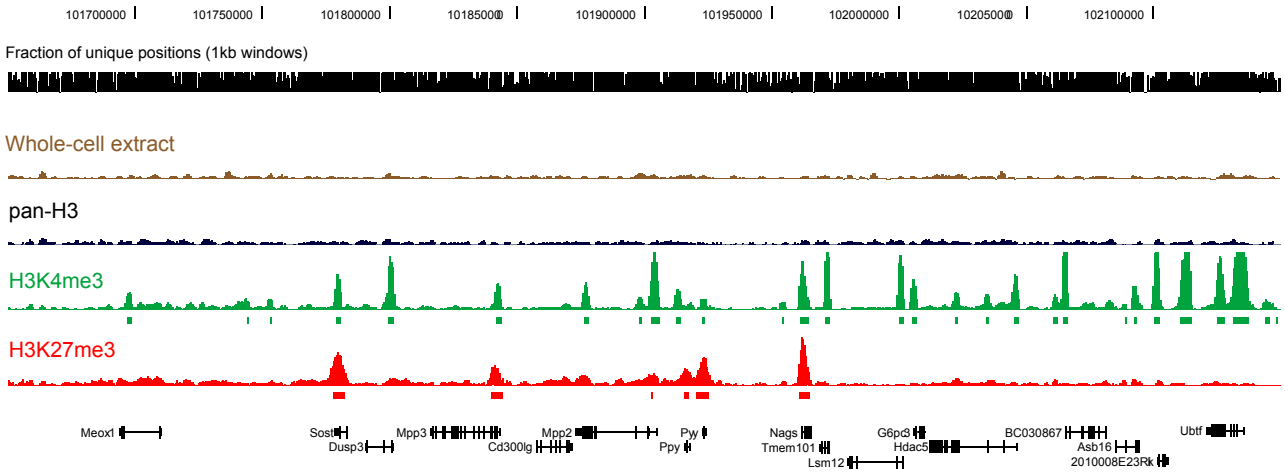
Moreover, the specificity of SMS read alignments is also high in practice: When reads from individual BAC clones are mapped onto the whole genome using our pipeline, >98% of mappable reads are placed correctly (at k=27, d=2). This implies that ChIP-Seq can accurately interrogate ~70% of the mammalian genome. By comparison, ChIP-chip yields data for at most ~50% because most repeats are ignored due to the problem of cross-hybridization.

Supplementary Figure 1



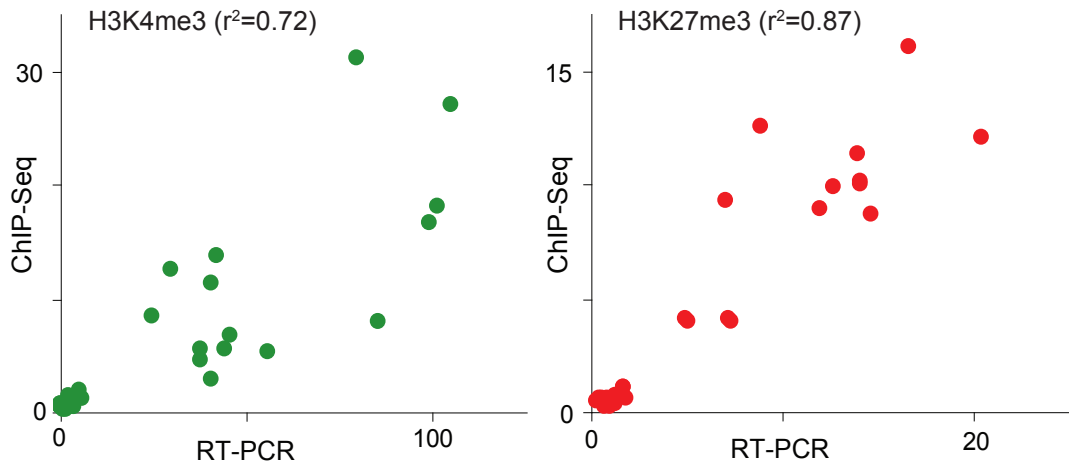
Cumulative distributions of fragment densities (averaged over 1-kb windows) across the mouse genome are shown for ES cell ChIPs of pan-H3 (blue), H3K4me3 (green) and H3K27me3 (red), and for unenriched whole-cell extracts (brown). The black curves show the distributions obtained from randomized placements of the of the same reads. The observed distributions for pan-H3 ChIP and whole-cell extract are virtually identical to the randomized distributions, indicating that ChIP-Seq generates unbiased data from unenriched samples. In contrast, the observed distributions for H3K4me3 and H3K27me3 enriched samples show clear excess of extreme values.

Supplementary Figure 2



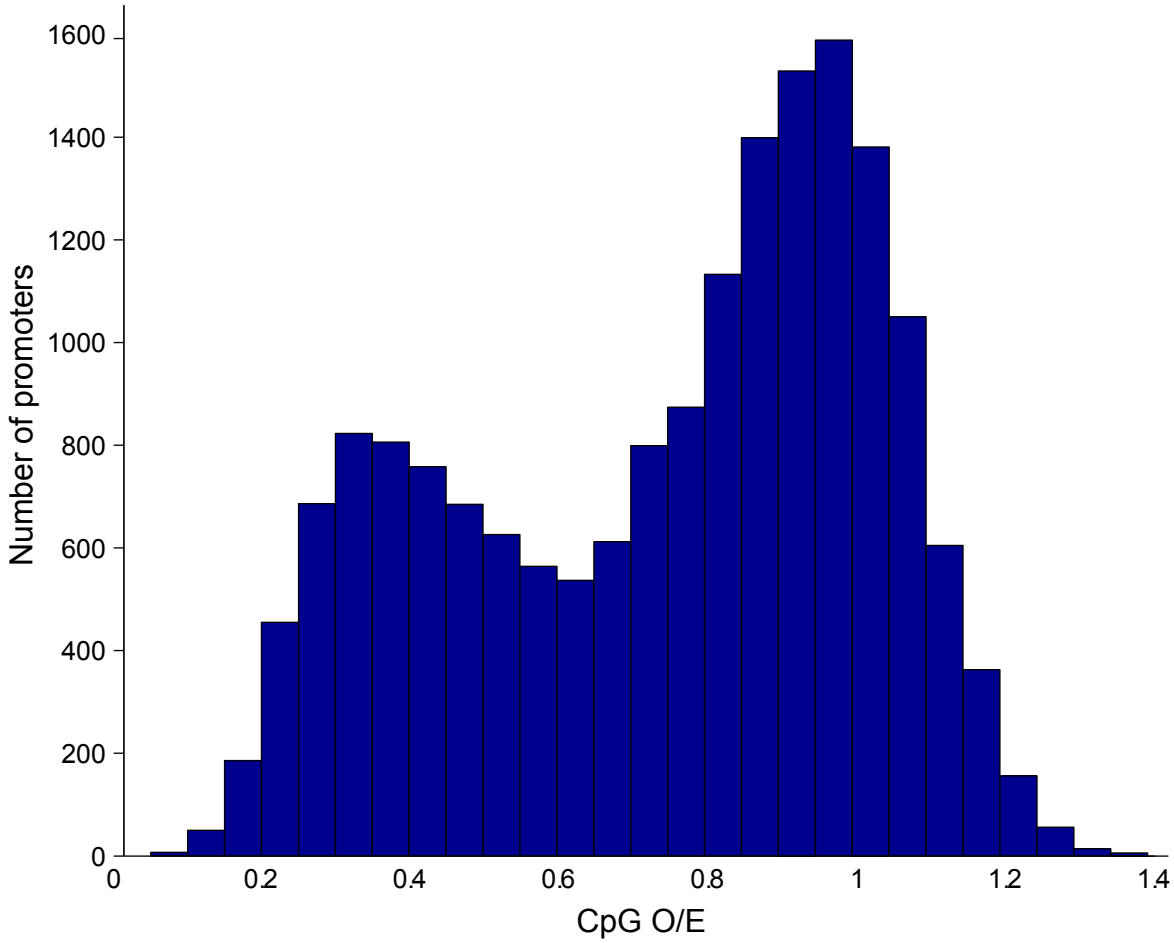
A representative comparison of ChIP-Seq fragment densities across a 500 kb interval on mouse chromosome 1. Rectangles beneath each density plot indicate significantly enriched intervals at the $p < 10e\text{-}5$ threshold (see Methods). Black bars at the top indicate the fraction of unique positions within 1kb windows at which ChIP-Seq reads can be uniquely aligned (at $k = 27$, $d = 2$; see Methods).

Supplementary Figure 3



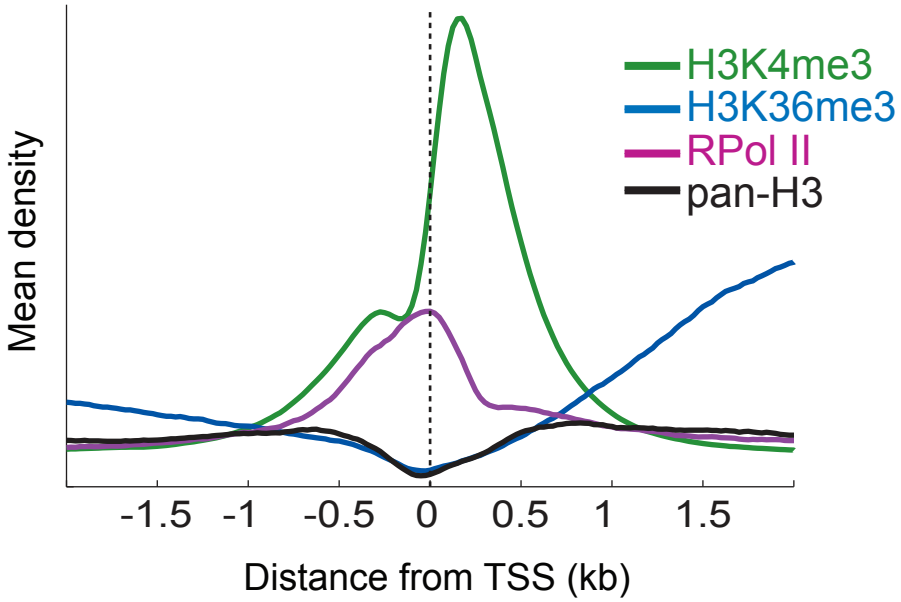
ChIP-Seq fragment densities (y-axis) are plotted against RT-PCR fold-enrichment (x-axis) for H3K4me3 (green) and H3K27me3 (red) at 60 selected sites in mouse ES cells. Notably 28 out of 29 sites (97%) identified as significantly enriched for one of the two modifications by ChIP-Seq were clearly differentiated from unenriched sites by RT-PCR, and 31 of 31 sites (100%) with no ChIP-Seq enrichment had no RT-PCR enrichment either.

Supplementary Figure 4



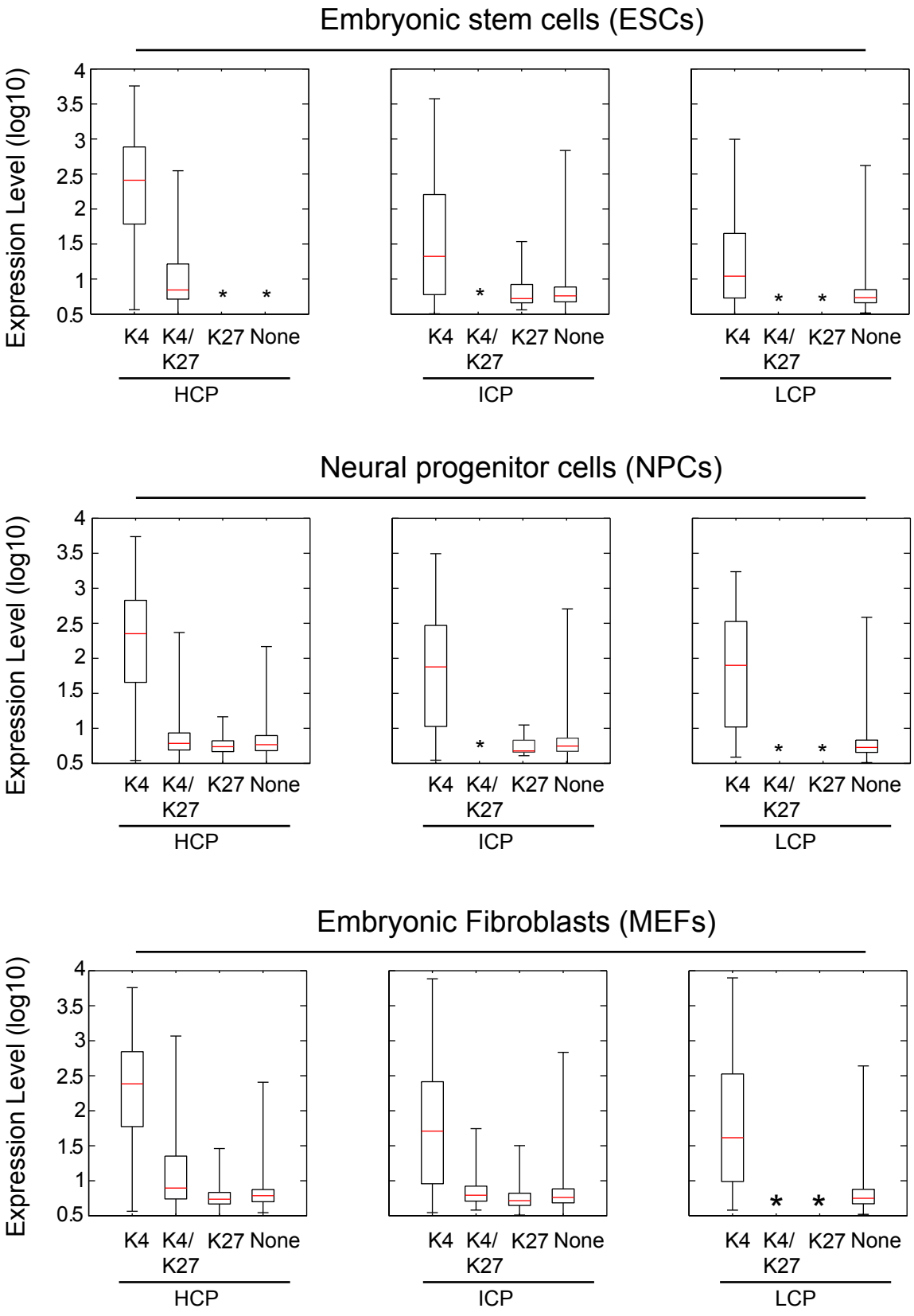
Histogram of the maximal observed to expected ratio of CpGs across 500 bp sliding windows from -0.5 kb to +2 kb at each analyzed promoter. The bimodal distribution reflects the two major classes of mammalian RNA polymerase II promoters.

Supplementary Figure 5



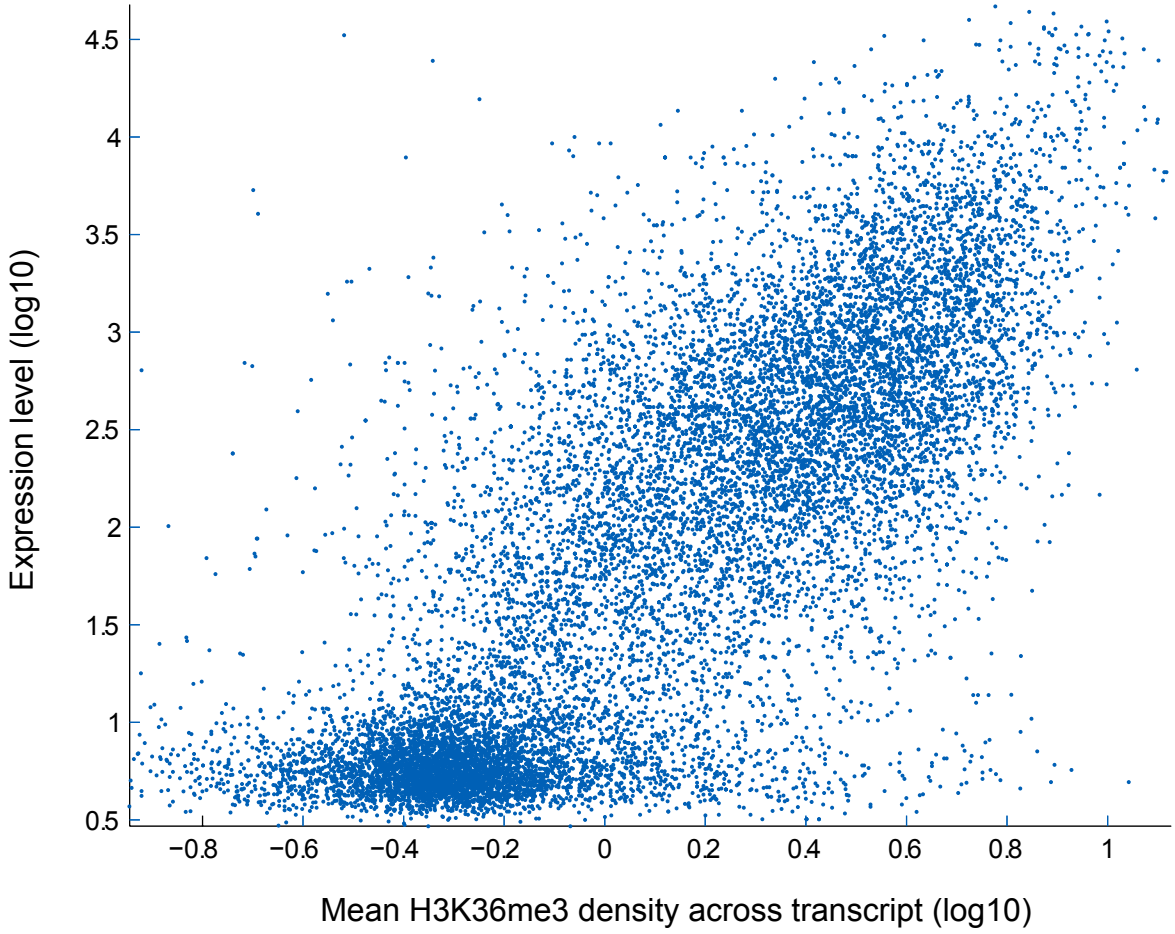
Composite profile of HCP promoters. Plots show mean ChIP-Seq fragment densities (scaled for comparison) of H3K4me3, H3K36me3, pan-H3 and RNA Polymerase II ChIP-Seq fragments over all analyzed high-CpG promoters. H3K4me3 marks a punctate interval peaking just downstream of the H3-depleted transcription start site, which is occupied by RNA Polymerase II. H3K36me3 marks begin roughly where H3K4me3 ends. H3K4me3 and H3K36me3 increasing in the negative direction likely represent bidirectional promoter activity.

Supplementary Figure 6



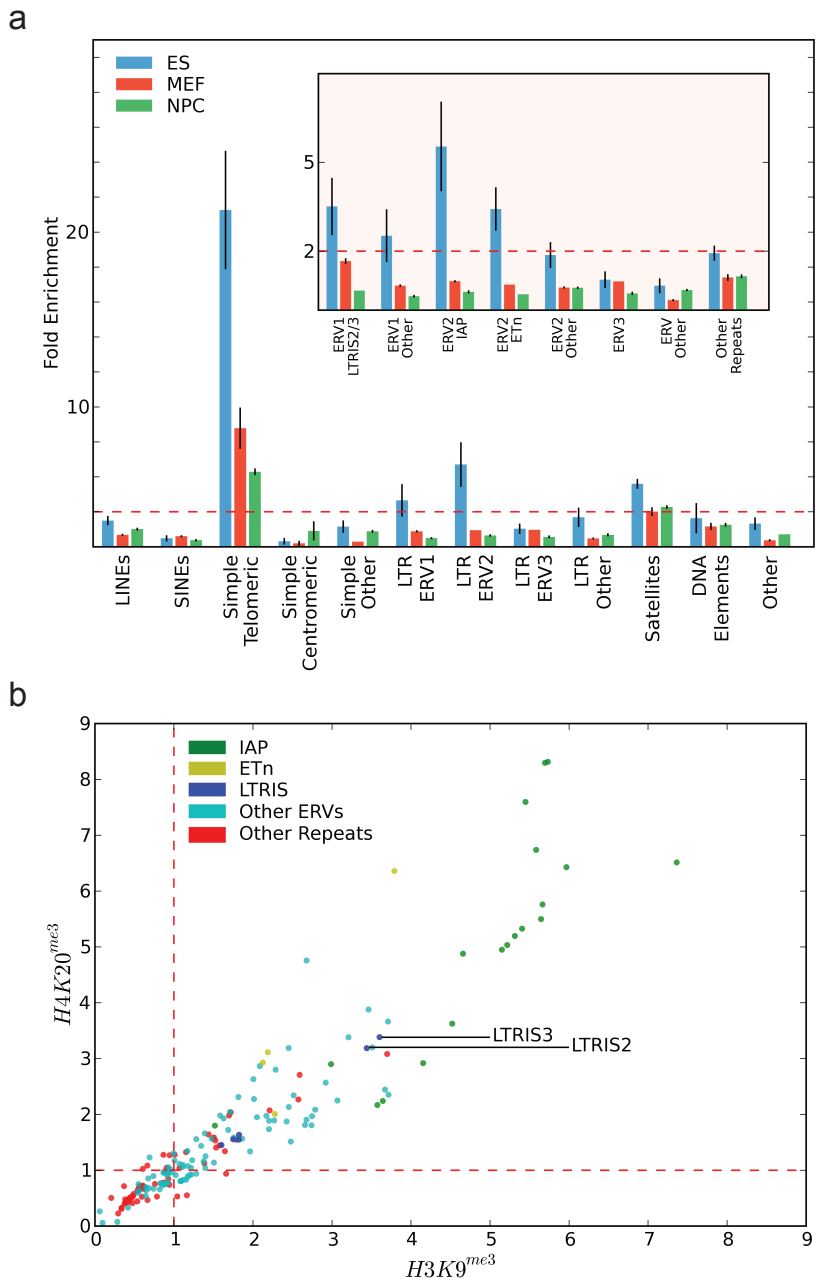
Boxplots showing the distributions of expression levels for genes in ES cells, NPCs and MEFs, according to promoter class and state. Red bar is median; box shows 25th and 75th percentiles; whiskers show 2.5th and 97.5th percentiles. Asterisks indicate class/state combinations with less than 15 genes.

Supplementary Figure 7



Scatter plot of H3K36me3 density across transcripts versus their expression levels as measured by Affymetrix GeneChips. The lower left-hand cluster corresponds to largely inactive genes.The range of H3K36me3 densities across most actively expressed genes spans ~1 order of magnitude, compared to ~3 for expression levels.

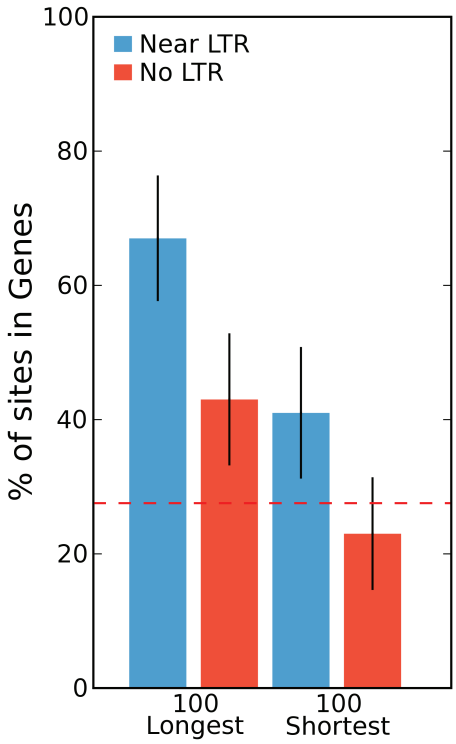
Supplementary Figure 8



(a) Simple telomeric repeats, satellite repeats, and class II endogenous retroviruses (LTR ERV2) all show significant enrichment for H3K9me3 in ES cells (blue). A weaker signal is seen for class I endogenous retroviruses (LTR ERV1). Both ERV1 and ERV2 elements lose the H3K9me3 marking in MEFs (red) and NPCs (green). The dashed line indicates twofold enrichment; values below 1 indicate depletion. Error bars show the difference in signal observed between sample runs. Inset: Intracisternal A particles (ERV2 IAP) and Early Transposon associated elements (ERV2 ETn), both known to be active in mice, are largely responsible for the enrichment of H3K9me3 observed for ERV2s in ES cells. Select families of ERV1s exhibit H3K9me3, including some members of the LTRIS family of repeats.

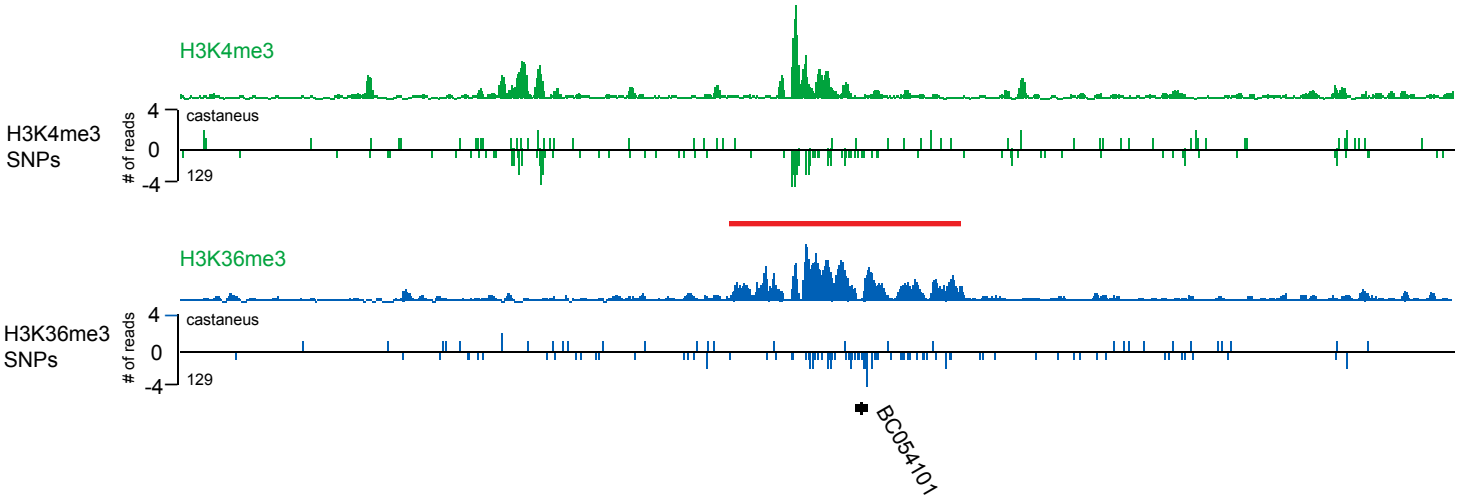
(b) H3K9me3 and H4K20me3 exhibit similar distribution at repeats, and are strongly enriched in active ERVs. ETn (yellow) and particularly IAP (green) elements, both known to be active ERVs in mice, exhibit strong signals. Several other ERV families are enriched for both markers, which may reflect continued activity. Several members of the little-studied LTRIS family (blue) fall into this category.

Supplementary Figure 9



Long H3K9me3 sites without nearby LTRs tend to be in genes. Of the H3K9me3 sites lacking nearby LTRs (blue), two-thirds of the 100 longest sites lie in genes, as compared to the 27.5% that would be expected for random alignable regions (dashed line). The 100 shortest sites lacking nearby LTRs show only a mild enrichment, illustrating the dependence of this effect on site length. H3K9me3 sites with nearby LTRs (red) are less likely to lie in genes. Error bars illustrate 95% confidence values.

Supplementary Figure 10



A transcribed region that exhibits allelic bias in the ChIP-Seq data from 129/*castaneus* hybrid ES cells is a candidate for imprinted or strain-specific expression. An interval of H3K36me3 enrichment (red bar) overlapping the transcript of unknown function BC054101. Of the 69 aligned reads within the enriched interval at the center of the locus, 64 were classified as 129 (maternal). Of the 61 aligned H3K4me3 reads in the same interval, 59 were classified as 129. This suggests near exclusive maternal transcription.