Cell, Volume *130*

Supplemental Data

A Chromatin Landmark and

Transcription Initiation at

Most Promoters in Human Cells

Matthew G. Guenther, Stuart S. Levine, Laurie A. Boyer, Rudolf Jaenisch, and Richard A. Young

Supplemental Experimental Procedures

Growth Conditions for Human Embryonic Stem Cells

Human embryonic stem (ES) cells were obtained from WiCell (Madison, WI; NIH Code WA09) and grown as described (Cowan et al., 2004). Briefly, passage 34 cells were grown in KO-DMEM medium supplemented with serum replacement, basic fibroblast growth factor (FGF), recombinant human leukemia inhibitory factor (LIF) and a human plasma protein fraction. Detailed protocol information on human ES cell growth conditions and culture reagents are available at http://www.mcb.harvard.edu/melton/hues.

In order to minimize any MEF contribution to our analysis, H9 cells were cultured on a low density of irradiated murine embryonic fibroblasts (ICR MEFs) resulting in a ratio of approximately >8:1 H9 cell to MEF (Figure S1). The culture of H9 on low-density MEFs had no adverse affects on cell morphology, growth rate, or undifferentiated status as determined by immunohistochemistry for pluripotency markers (e.g. Oct4, SSEA-3, Tra-1-60; see below). In addition, H9 cells grown on a minimal feeder layer maintained the ability to generate derivates of ectoderm, mesoderm, and endoderm upon differentiation (see below).

Quality Control for Human Embryonic Stem Cells

Immunohistochemical Analysis of Pluripotency Markers

For analysis of pluripotency markers, cells were fixed in 4% paraformaldehyde for 30 minutes at room temperature and incubated overnight at 4°C in blocking solution (5 ml Normal Donkey Solution:195 ml PBS + 0.1% Triton-X) (Figure S2). After a brief wash in PBS, cells were incubated with primary antibodies to Oct-3/4 (Santa Cruz sc-9081), SSEA-3 (MC-631) (Solter and Knowles, 1979), SSEA-4 (MC-813-70) (Solter and Knowles, 1979), Tra-1-60 (MAB4360; Chemicon International), and Tra-1-81 (MAB4381; Chemicon International) in blocking solution overnight at 4°C. Following incubation with primary antibody, cells were incubated with either rhodamine red or FITC-conjugated secondary antibody (Jackson Labs) for 2-5hrs at 4°C. Nuclei were stained with 4',6-diamidine-2-phenylidole dihydrochloride (DAPI). Epifluorescent images were obtained using a fluorescent microscope (Nikon TE300). Data is shown for Oct4 and SSEA-3. Our analysis indicated that >90% of the H9 cells were strongly positive for all pluripotency markers.

Alkaline phosphatase activity of human ES cells was analyzed using the Vector Red Alkaline Phosphatase Substrate Kit (Cat. No. SK-5100; Vector Laboratories) according to manufacturer's specifications and the reaction product was visualized using fluorescent microscopy.

Teratoma Formation
Teratomas were induced by injecting 2-5 x 10^6 cells into the subcutaneous tissue above the rear haunch of 6 week old Nude Swiss (athymic, immunocompromised) mice. Eight to twelve weeks post-injection, teratomas were harvested and fixed overnight in 4% paraformaldehyde at 4°C. Samples were then immersed in 30% sucrose overnight before embedding the tissue in O.C.T freezing compound (Tissue-Tek). Cryosections were obtained and 10 μm sections were incubated with the appropriate antibodies as above and analyzed for the presence of the following differentiation markers by confocal microscopy (LSM 210): neuronal class II β-tubulin, Tuj1 (ectoderm; MMS-435P Covance); striated muscle-specific myosin, MF20 (mesoderm; kind gift from D. Fischman), and alphafetoprotein (endoderm; DAKO) (Figure S3). Nuclei were stained blue with 4',6-diamidine-2-phenylidole dihydrochloride (DAPI). Antibody reactivity was detected for markers of all three germ layers confirming that the human embryonic cells used in our analysis had maintained differentiation potential.

Embryoid Bodies (EB)
ES cells were harvested by enzymatic digestion and EBs were allowed to form by plating ~1 X 10^6 cells/well in suspension on 6-well non-adherent, low cluster dishes for 30 days. EBs were grown in the absence of leukemia inhibitory factor (LIF) and basic fibroblast growth factor (FGF) in culture medium containing 2x serum replacement. EBs were then harvested, fixed for 30 minutes in 4% paraformaldehyde at room temperature, and placed in 30% sucrose overnight prior to embedding the tissue in O.C.T. freezing compound (Tissue-Tek). Cryosections were obtained as described for teratoma formation. Confocal images were obtained for all three germ layer markers again confirming that the H9 cells used in our analysis have maintained differentiation potential (data not shown; results similar to those shown in Figure S3).

Antibodies and Controls
H3K4me3-bound genomic DNA was isolated from whole cell lysate using an epitope specific rabbit polyclonal antibody purchased from Abcam (AB8580) (Santos-Rosa et al., 2002). Chromatin immunoprecipitations against H3K4me3 were compared to reference DNA obtained by chromatin immunoprecipitation of total histone H3 (Abcam AB1791; epitope derived from C-terminal 100 amino acids of histone H3) to normalize for nucleosome density. A representative scatter plot of H3K4me3 compared to total histone H3 and of histone H3 compared to genomic DNA can be seen in Figure S4. A summary of genes bound with high confidence for this and the antibodies below is listed in Table S2.

RNA polymerase II initiation form-bound genomic DNA was isolated from whole cell lysate using 8WG16, a mouse monoclonal antibody. This antibody preferentially binds a form of RNA polymerase II that lacks phosphorylation at the C-terminal domain of the largest subunit of polymerase (Patturajan et al., 1999; Cho et al., 2001; Jones et al., 2004) although this preference can be subject to experimental conditions. Elongating RNA polymerase II (phosphoserine-5) bound genomic DNA was isolated from whole cell lysate using 4H8, a mouse monoclonal antibody (Covance MMS1-28P, (Kristjuhan et al., 2002; Brodsky et al., 2005).

H3K9,14Ac-bound genomic DNA was isolated from hES whole cell lysate using Upstate antibody 06-599, a rabbit polyclonal antibody. This antibody recognizes both the lysine-9 and lysine-14 acetylated residues on histone H3. Chromatin immunoprecipitations against H3K9,14Ac were compared to reference DNA obtained by chromatin immunoprecipitation of total histone H3 (Abcam AB1791; epitope derived from C-terminal 100 amino acids of histone H3) to normalize for nucleosome density.

H3K36me3-bound genomic DNA was isolated from hES whole cell lysate using rabbit polyclonal antibody purchased from Abcam (AB9050). Chromatin immunoprecipitations against H3K36me3 were compared to reference DNA obtained by chromatin immunoprecipitation of

total histone H3 (Abcam AB1791; epitope derived from C-terminal 100 amino acids of histone H3) to normalize for nucleosome density.

H3K79me2-bound genomic DNA was isolated from hES whole cell lysate using Abcam antibody AB3594. Chromatin immunoprecipitations against H3K79me2 were compared to reference DNA obtained by chromatin immunoprecipitation of total histone H3 (Abcam AB1791; epitope derived from C-terminal 100 amino acids of histone H3) to normalize for nucleosome density.

IgG control immunoprecipitaitons were preformed from hES whole cell lysate using goat IgG from Santa Cruz Biotechnology (sc-2043). A representative scatter plot of IgG is shown in Figure S4. Note the very minimal enrichment of oligos compared to H3K4me3.The few oligos that appear enriched are randomly distributed (data not shown).

E2F4 immunoprecipitations were preformed from hES whole cell lysate using antibodies purchased from Santa Cruz Biotech (sc-1082) (Boyer et al., 2005). This antibody has been shown to specifically recognize previously reported E2F4 target genes (Ren et al., 2002; Weinmann et al., 2002).

Normalizing levels of histone modification to experimentally derived levels of histone is essential for accurate interpretation of the data. For example, if the signal for histone H3 acetylation is reduced at a locus, and we lack information on the level of histones at that locus, we cannot know whether the loss of H3Ac signal is due to 1) reduced acetylation of nucleosomes or 2) normal acetylation of fewer nucleosomes. Additionally, if an increase in ratio of modified histone to total histone was primarily due to the loss of total histone, we would expect the composite profiles of each histone modifications to occur directly over minimum histone levels. Instead, we observe substantial displacement of these peaks relative to the total H3 minima (Figure S8). This supports both the specificity of the antibodies, and the primary role of the histone modification in the binding ratio.

To confirm the the primary role of the histone modification in the binding ratio, we performed additional ChIP-chip experiments to more closely examine H3K4me3 levels across known transcription start sites relative to total DNA content. This allows us to discriminate between the possibilities that either 1) the relative increase in H3K4me3-modified nucleosomes is due to a removal of unmethylated histones or 2) the increased methylation of H3K4 is due to targeted methylation by H3K4 methyltransferases. If the presence of H3K4me3 binding events were due to a diminishment/absence of histones (due to selective removal of non-methylated histones or otherwise), then the H3K4me3 normalized to total DNA content would not show enrichment at these regions. The results of these experiments are presented in figure S10. We find that the levels of H3K4me3 enrichment are largely unchanged with the exception of a small region at the transcription start site, which reflects the decrease in histone content at the promoter. The contribution to the probe by probe enrichment therefore varies by position, but at the transcription start site, it is 9%. Overall, we find that 95% of our high-confidence H3K4me3 targets are preserved when we normalize to total DNA content. While it is difficult to determine the extent to which removal of non-methylated nucleosomes might account for the enrichment of H3K4me3-modified nucleosomes at promoters, the enrichment of H3K4me3 even against genomic DNA suggests that it is unlikely to be the dominant mechanism.

Chromatin Immunoprecipitation

Protocols describing all materials and methods have been previously described and can be downloaded from http://web.wi.mit.edu/young/hES_PRC.

We performed independent immunoprecipitations for each analysis. Human WA09 embryonic stem cells were grown to a final count of $5x10^7$ – $1x10^8$ cells for each location analysis experiment. Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 15 minutes at room temperature. Cells were rinsed twice

with 1xPBS and harvested using a silicon scraper and flash frozen in liquid nitrogen. Cells were stored at –80^{o}C prior to use.

Cells were resuspended, lysed in lysis buffers and sonicated to solubilize and shear crosslinked DNA. Sonication conditions vary depending on cells, culture conditions, crosslinking and equipment. We used a Misonix Sonicator 3000 and sonicated at power 7 for 10 x 30 second pulses (90 second pause between pulses). Samples were kept on ice at all times.

The resulting whole cell extract was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with approximately 10 µg of the appropriate antibody. The immunoprecipitation was allowed to proceed overnight.

Beads were washed 5 times with RIPA buffer and 1 time with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65°C with occasional vortexing and crosslinking was reversed by overnight incubation at 65°C. Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal.

Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and multiple phenol:chloroform:isoamyl alcohol extractions. Purified DNA was blunted and ligated to linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled and purified using Bioprime random primer labeling kits (Invitrogen, immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA was labeled with Cy3 fluorophore).

Labeled DNA was mixed (5-6 µg each of immunoenriched and whole cell extract DNA) and hybridized to arrays in Agilent hybridization chambers for up to 40 hours at 40°C. Arrays were then washed and scanned.

Slides were scanned using an Agilent DNA microarray scanner BA. PMT settings were set manually to normalize bulk signal in the Cy3 and Cy5 channel. For efficient batch processing of scans, we used either Genepix software (Axon) or Feature Extractor (Agilent). Scans were automatically aligned and then manually examined for abnormal features. Intensity data were then extracted in batch.

Slide Reuse

Reuse of Agilent slides for ChIP-chip experiments has been described previously (Pokholok et al., 2005; Pokholok et al., 2006). Briefly, arrays were washed in acetonitrile for 3 min and then rinsed in room temperature stripping buffer (100mM Potassium phosphate, pH6.6) for 1 min. Slides were transferred to 65^{0}C stripping buffer bath and brought to 100^{0}C. After boiling at 100^{0}C for 3-5 min, arrays were transferred to room temperature stripping buffer and allowed to cool for 2 min. Arrays were then dried and hybridized as normal. Whole genome slides were used a total of 3 times. First use arrays were hybridized to hES H3K4me3 immunoenriched material, stripped and hybridized to primary hepatocyte H3K4me3 immunoenriched material (second use). These slides were then stripped and hybridized to REH (B cell) H3K4me3 immunoenriched material (third use). All results were verified independently by hybridization to first use human promoter arrays (below).

Array Design

Whole-Genome Array

The human genome array was purchased from Agilent Technology (www.agilent.com). The array consists of 25 slides each containing ~244,000 60mer oligos (slide ID 14818-14841 and 14843) covering the entire non-repeated portion of the human genome at a density of about 1 oligo per 250bp.

185K Promoter Array
The human promoter array was purchased from Agilent Technology (www.agilent.com). The array consists of 2 slides each containing ~185,000 60mer oligos (slide ID 14154, 14155) designed to cover regions between approximately -5.5 kb and +2.5 kb relative to the transcription start sites of ~17,350 genes at a density of about 1 oligo per 250 bp.

44k Promoter Array
The human promoter array was purchased from Agilent Technology (www.agilent.com). The array consists of 10 slides each containing ~44,000 60mer oligos designed to cover regions between approximately -8kb and +2kb relative to the transcription start site. The design of these arrays are discussed in detail elsewhere (Boyer et al., 2005).

Data Normalization and Analysis
We used both Feature extractor (Agilent) and GenePix software (Axon) to obtain background-subtracted intensity values for each fluorophore for every feature on the whole genome and sub genomic arrays respectively. Among the Agilent controls is a set of negative control spots that contain 60-mer sequences that do not cross-hybridize to human genomic DNA. We calculated the median intensity of these negative control spots in each channel and then subtracted this number from the intensities of all other features.

To correct for different amounts of each sample of DNA hybridized to the chip, the negative control-subtracted median intensity value of control oligonucleotides from the Cy3-enriched DNA channel was then divided by the median of the control oligonucleotides from the Cy5-enriched DNA channel. This yielded a normalization factor that was applied to each intensity in the Cy5 DNA channel.

Next, we calculated the log of the ratio of intensity in theCy3-enriched channel to intensity in the Cy5 channel for each probe and used a whole chip error model (Hughes et al., 2000) to calculate confidence values for each spot on each array (single probe p-value). This error model functions by converting the intensity information in both channels to an X score which is dependent on both the absolute value of intensities and background noise in each channel using an f-score calculated as described (Boyer et al., 2005) for promoter regions or using a score of 0.3 for tiled arrays. When available, replicate data were combined, using the X scores and ratios of individual replicates to weight each replicate's contribution to a combined X score and ratio. The X scores for the combined replicate are assumed to be normally distributed which allows for calculation of a p-value for the enrichment ratio seen at each feature. P-values were also calculated based on a second model assuming that, for any range of signal intensities, IP:control ratios below 1 represent noise (as the immunoprecipitation should only result in enrichment of specific signals) and the distribution of noise among ratios above 1 is the reflection of the distribution of noise among ratios below 1.

High-Confidence Enrichment
To automatically determine bound regions in the datasets, we developed an algorithm to incorporate information from neighboring probes. For each 60-mer, we calculated the average X score of the 60-mer and its two immediate neighbors. If a feature was flagged as abnormal during scanning, we assumed it gave a neutral contribution to the average X score. Similarly, if an adjacent feature was beyond a reasonable distance from the probe (1000 bp), we assumed it gave a neutral contribution to the average X score. The distance threshold of 1000 bp was determined based on the maximum size of labeled DNA fragments put into the hybridization. Since the maximum fragment size was approximately 550 bp, we reasoned that probes separated by 1000 or more bp would not be able to contribute reliable information about a binding event halfway between them.

This set of averaged values gave us a new distribution that was subsequently used to calculate p-values of average X (probe set p-values). If the probe set p-value was less than 0.001, the three probes were marked as potentially bound.

As most probes were spaced within the resolution limit of chromatin immunoprecipitation, we next required that multiple probes in the probe set provide evidence of a binding event. Candidate bound probe sets were required to pass one of two additional filters: two of the three probes in a probe set must each have single probe p-values < 0.005 or the center probe in the probe set has a single probe p-value < 0.001 and one of the flanking probes has a single point p-value < 0.1. These two filters cover situations where a binding event occurs midway between two probes and each weakly detects the event or where a binding event occurs very close to one probe and is very weakly detected by a neighboring probe. Individual probe sets that passed these criteria and were spaced closely together were collapsed into bound regions if the center probes of the probe sets were within 1000 bp of each other.

Simple Ratio Enrichment

To detect regions of the genome that show lower levels of enrichment, we also utilized a simple ratio threshold for binding. This used a modification of the algorithm described above and required only that the weighted average of the median of ratios for the center probe of a probe set was greater then 2.0. These probe sets were then collapsed into bound regions as above.

Graphical display of enrichment ratios at individual regions were derived from unprocessed enrichment ratios for all probes within a genomic region (ChIP versus whole genomic DNA for Pol II ; ChIP versus total H3 for histone modifications). Each individual probe was graphed as a sliding average value of the individual probe averaged with the nearest 5' probe and the nearest 3' probe (sliding average of 3 probes).

Replicate Information in Subsets of the Human Genome

To confirm the accuracy of the genome-wide enrichment of H3K4me3, two biological replicates for each sample were hybridized to Agilent 185k promoter arrays. Table S4 shows the lists of genes enriched for H3K4me3 in hES cells, Liver, and REH cells. In all cases, over 98% of the genes enriched on the promoter proximal arrays were also enriched on the whole genome arrays. H3K4me3 enrichment on whole genome arrays was found at an additional 700-1500 genes (5-15%), reflecting both slightly stronger enrichment ratios, and the presence of enriched alternate start sites tiled only on the whole genome arrays.

Comparing Enriched Regions to Known and Predicted Genes

The coordinates for the complete lists of H3K4me3 enriched sites can be found in Table S1, Table S7, and Table S8 for embryonic stem cells, liver samples, and REH cells respectively.

Comparisons to Known Genes

Enriched regions were compared relative to transcript start and stop coordinates of known genes compiled from four different databases: RefSeq (Pruitt et al., 2005), Mammalian Gene Collection (MGC) (Gerhard et al., 2004), Ensembl (Hubbard et al., 2005), and University of California Santa Cruz (UCSC) Known Genes (genome.ucsc.edu)(Kent et al., 2002). All coordinate information was downloaded in January 2005 from the UCSC Genome Browser (NCBI build 35). Of the 19,360 H3K4me3 enriched regions in human embryonic stem cells, 13,299 (67%) occurred within 1 kb of gene starts from one of these 4 databases (Table S1)

To convert bound transcription start sites to more useful gene names, we used conversion tables downloaded from UCSC and Ensembl to automatically assign EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/) gene IDs and symbols to the RefSeq, MGC, Ensembl, UCSC Known Gene. This resulted in a total of 12,820 EntrezGene genes being highly enriched for H3K4me3 in human ES Cells (Table S2).

Fraction of Transcription Start Sites Enriched for H3K4me3in ES Cells
We used several human gene databases to identify the fraction of annotated transcription start sites enriched for H3K4me3 in hES cells. For each database, we calculated the percentage of annotated transcription start sites that lie within 1 kb of an enriched region (MGC 78%, RefSeq 74%, Ensembl 52%, UCSC Known Genes 69%).

Comparisons to Predicted Genes
The locations of H3K4me3 bound regions were also compared relative to transcript start and stop coordinates of predicted genes compiled from eight different databases; GenScan (Burge and Karlin, 1997), GeneID (Parra et al., 2000), FirstEF (Davuluri et al., 2001), ACEview (www.aceview.org), ECgene (Kim et al., 2005), UniGene (www.ncbi.nlm.nih.gov/UniGene), UCSC RetroFinder (Morillon et al., 2003) and Non-human mRNAs (Kent et al., 2002). These gene models are generally derived through *ab initio* computational gene modeling (GenScan, GeneID and FirstEF) or EST clustering and alignment to the human genome (ACEview, ECgene, UniGene, UCSC RetroFinder and Non-human mRNAs). All predictions were derived from downloads of coordinates of predicted human genes mapped to NCBI build 35 of the public human genome sequence from UCSC in January 2005. Of the 6061 H3K4me3 enriched not mapping to a known gene, 4741 mapped within 1 kb of the start site of either FirstEF, ACEview, or UniGene transcripts (Table S1). Therefore, a total of 18040 (93%) H3K4me3 enriched regions corresponded to a known or predicted transcription start site. Importantly, because nearly all H3K4me3 sites are proximal to known or predicted start sites, this suggests that as many as one-third of transcription initiation sites remain to be annotated.

Thresholds and the Fraction of Genes Called
For simplicity in understanding ChIP/chip data, genes have been assigned as either bound or unbound based on a single threshold model. However, binding data is rarely bimodal and instead represents a continuum that reflects the frequency of occupancy of a particular binding site in a large population of cells and is affected by the availability of the epitope, the quality of the oligos, shear distribution, and the density of target sites. In previous studies we have identified thresholds that minimize false positives on genome-wide analyses while accepting fairly high false negative rates. This threshold is termed “high confidence” and gene assignments based on this cutoff are described in Table S2. For H3K4me3 we find 74%, 71%, and 64% of genes bound at high confidence in hES, hepatocytes and B cells respectively. If we assume a 30% false negative rate and 4% false positive rate (Lee et al., 2006), the expected number of genes modified by H3K4me3 would be 79%, 77%, and 73% respectively.

To identify additional genes that have a high probability of being enriched, we used a simpler methodology requiring only a high enrichment ratio (>2 fold). This identified an addition 750 genes in hES cells as enriched while failing to identify only 21 called as high confidence (out of over 12,000). These genes cover a broad range of proteins and include portions of the HoxB and HoxD loci that were previously identified as enriched for H3K4me3 (Bernstein et al., 2006). These genes may have been missed in the more stringent error model because of poor probe placement or weak intensities associated with the promoter proximal oligonucleotides.

Human Expression Data
Detection of Transcripts from ES Cells
We collected 7 previously published ES cell expression datasets for comparison with our binding data (Lee et al., 2004). The expression data, gathered using massively parallel signature sequencing (MPSS) and Affymetrix gene expression arrays (Table S3), were processed as follows:

MPSS data: Three MPSS datasets were collected, two from a pool of the ES cell lines H1, H7 and H9 (Brandenberger et al., 2004; Wei et al., 2005) and one for HES-2 (Wei et al., 2005). For each study, only MPSS tags detected at or over 4 transcripts per million (tpm) were used. In addition, the data provided by Wei and colleagues (Wei et al., 2005) allowed us to select only those tags that could be mapped to a single unique location in the human genome. For tags without a corresponding EntrezGene ID, IDs were assigned using the gene name or RNA accession numbers provided by the authors. Genes designated as "detected" were identified in all 3 pools while only those absent from all 3 pools were considered "not detected".

Gene expression microarray data: Four Affymetrix HG-U133 gene expression datasets were collected for the cell lines H1 (Sato et al., 2003) and H9 (Abeyta et al., 2004). Each cell line was analyzed by the authors in triplicate. EntrezGene IDs were assigned to the probe-sets using Affymetrix annotation or using RNA accession numbers provides by the authors. For each probe-set, we counted the number of "Present" calls in the three replicate array experiments performed for each cell line. Most genes are represented by more than one probe-set and, to enable comparison to MPSS and RNA polymerase II binding data, we then found the maximum number of P calls for each gene (defined by unique EntrezGene ID). A gene was defined as detected if it was called "Present" in all 3 replicate arrays and absent if it was not detected in any of the replicate arrays.

ES Cell Expression Relative to Differentiated Cell Types

In order to compare ES cells with as many human cell and tissue types as possible, we combined the data from three studies, all performed using the Affymetrix HG-U133A platform: 3 replicates of H1 ES cells (Sato et al., 2003), 3 replicates each of H9, HSF1 and HSF6 ES cells (Abeyta et al., 2004) and 2 replicates of 79 other human cell and tissue types (Su et al., 2004). We extracted data from the original CEL files from each array and scaled the data to a median signal of 150 in GCOS (Affymetrix). We then exported the data, created expression ratios using the median gene expression of each gene across all arrays, transformed the data into log base2 and median centered both gene and arrays (so that the median log2 expression ratio for each gene and each array is 0). EntrezGeneIDs were assigned to each probe-set and for genes with multiple probe-sets, the expression ratios averaged. This resulted in a set of 12,968 unique genes. The genes were then sorted based on the relative expression of the H1 ES cells (which had the closest agreement with our previous studies (Lee et al., 2006) and the top 10%, center 10% and bottom 10% were selected for creating composite profiles.

While these lists of detected and not detected genes provided a series of very high-confidence targets for comparing ChIP-chip data, the lists exclude nearly half of the genome. To obtain a picture of the overall levels of H3K4 methylated nucleosomes at inactive genes, we used a simple voting mechanism to assign every gene as either active or inactive by calling a gene active if one Affymetrix probe for a given gene was detected in at least half of the experiments we analyzed (Figure S7). Using this algorithm, ~50% of genes were considered active, over 90% of which had H3K4me3-modified nucleosomes at one of their promoters. Among the remaining half of the genes, 60% had H3K4me3-modified nucleosomes at the transcription start site, leaving 20% of the total population where the promoters did not contain H3K4me3-modified nucleosomes.

RT-PCR Analysis

Three independent standard curves were produced for three independent cDNAs (ANG, GPR143, KCNJ3) contained in pCMV6-XL4 (OriGene Technologies). Total amounts of 100ng, 10ng, 1ng, 0.1ng, 10pg, 1pg, 0.1pg, 0.01pg, 0.001pg, .0001pg were subjected to quantitative real time PCR (described below) in duplicate. Actual molecules for each dilution curve were calculated based on the molecular weight of the cDNA. A standard curve was then derived for

each cDNA and then averaged to produce a composite standard curve (Ct versus molecules present) for which all test measurements were compared (Figure 4).

6 x 10^6 hES cells cultured as above were enriched from feeder MEFs by trypsinization. RNA was extracted from hES cells by TRIAZOL method and precipitation (Invitrogen). Total RNA was reverse transcribed (RT) by method of Invitrogen Superscript III First Strand Synthesis System using both oligo dT and random hexamer primers to produce cDNA. Total cDNA from 6 x 10^6 cells was contained in 100μl. 1μl cDNA (1/100) was used for each individual quantitative PCR measurement. cDNA was amplified using TaqMan Pre-developed gene expression assays (20X mixture supplied by Applied Biosystems which included pre-optimized primers and probe;Applied Biosystems). Duplicate reactions were performed in a total of 40μl using Taqman universal PCR master mix in an Applied Biosciences 7500 Real Time PCR Thermocycler. Detection of abundance was determined by measuring the point during cycling when amplification could first be detected, rather than the endpoint of the 40 cycle reaction. This cycle threshold (Ct) value corresponds to the fractional cycle number where the florescent Taqman probe increases above a fixed threshold determined by the ABI Prism 7000 Sequence Detection System software. The measured Ct value was used to calculate the estimated transcripts present in the test sample using relative quantitation to the composite standard curve. All standards, controls and samples were assayed according to the same standard criteria. Each gene determined to be active by Affymetrix and MPSS methods (RPLPO, HMGB2, CCNB1, EIF2C3) and genes determined to be inactive by Affymetrix and MPSS methods (HOP, GUCY1A3, KITLG, SOX9, PLD1, INCENP, ERCC4, CSF1, HOXB1, GPR143, DSC3, NFKBIL2, ANG, LEFTY1, LNPEP, TCF21, RAP1GA1, KLF2, KCNJ3, TNFSF7, PRRG2, FOXG1B, DLX5, KCNH1, ACADL, CXCL2, HOXD12, ONECUT1, LYST, PCSK2, AVPR1B, SMP3, EPHX2, UCP1, AGC1, ERAF, HBA1, MLNR) were measured using unique Taqman pre-developed gene expression assays for each species.

Detection of 5' RNA Transcripts at Inactive Genes

Our findings indicate that transcription initiation occurs at inactive genes that do not produce full-length transcripts. In order to determine if shorter 5' RNAs can be detected at these inactive genes, we assayed the presence of 5' RNAs at 10 genes that 1) were H3K4me3 modified at their promoters and 2) were found to be inactive by affymetrix expression analysis, MPSS analysis and TaqMan qRT-PCR analysis (<1 transcript per cell) (Figure 4, described above). If transcription initiation is occurring at these genes, we would expect to detect 5' RNAs. We also measured 5' RNA production at 2 active genes that produce full-length transcript and serve as positive controls.

Total RNA from 15 x 10^6 hES cells was obtained by TRIAZOL extraction and precipitation (Invitrogen). Total RNA was treated with DNAse (DNA free kit, Ambion) to remove any trace genomic DNA . Total RNA was reverse transcribed using Superscript RT kit (Invitrogen) using polyT, random hexamer, and specific primers (designed to create first strand PCR template from between 1 and 70 bp relative to transcription start site). Specific primers were designed against the region 30-70bp downstream of the transcription start site for each gene and also served as reverse primers for later qPCR analysis (below). Primers were designed against the following genes: (1)HMGB2, (2)CCNB1, (3)AVPR3, (4)ONECUT1, (5)HBA1, (6)ACADL, (7)ERAF, (8)DLX5, (9)AGC1, (10)CXCL2, (11)PSK2, (12)HOXD12, (13)KCNJ3, (14)OR9Q1. First strand synthesis mix was diluted to the RNA equivalent of 18,000 cells per ul and 0.5ul used per assay.

Pre-optimized TaqMan probes (Applied Biosystems) were first used to verify the absence of full-length transcript at 'inactive' genes. Duplicate reactions were performed in a total of 40μl using Taqman universal PCR master mix in an Applied Biosciences 7500 Real Time PCR Thermocycler. Detection of abundance was determined by measuring the point during cycling when amplification could first be detected (auto ct value determined by Applied Biosciences

7500 software). The measured Ct value was used to calculate the estimated transcripts present in the test sample using relative quantitation to the composite standard curve (above, figure S5).

5' RNA species were detected using 18-mer forward and reverse primers within the region 1-70bp relative to transcription start site for each gene. Duplicate reactions were performed in a total of 40μl using Power SYBR Green universal PCR master mix in an Applied Biosciences 7500 Real Time PCR Thermocycler using an extension/annealing temperature of 55^{o}C. Detection of abundance was determined by measuring the point during cycling when amplification could first be detected (auto ct value determined by Applied Biosciences 7500 software). The measured Ct value was used to calculate the estimated transcripts present in the test sample using relative quantitation to the composite standard curve (above, figure S5). Duplicate control reactions using RNA that was not reverse transcribed (no RT/no template control) were assayed as above using SYBR green assay conditions (Applied Biosystems). 5' Primer sequences used were:
(1)HMGB2 For-TGGCCTAGCTCGTCAAGTT Rev-CCTCAGCCTCTTGTTTTGC,
(2)CCNB1 For-GCCAATAAGGAGGGAGCA Rev-AGCCAGCCTAGCCTCAGA,
(3)AVPR3 For-GCTGCTCACCAGGCAGAG Rev-AGTCGGTGTCGCGCTCAG, (4)ONECUT1 For-GAGAGGAAGGAAGGCAACAG Rev-ACGGAGTCCGGTCTTCAC,
(5)HBA1 For-CTCTTCTGGTCCCCACAGA Rev-GGCAGGAGACAGCACCAT,
(6)ACADL For-CGCTTTTTTGGGAGGACA Rev-ATCAGCTGAGGCGTCCAC,
(7)ERAF For-TGGCACAGAGAGATTCACG Rev-TCTCACCCACACTCTTGAGG,
(8)DLX5 For-GGCCGGAGACAGAGACTTC Rev-CGAGGAGGAGACTGGGAGT, (9)AGC1 For-CCAGGTGTGTGGGACTGA Rev-TGGACTCCCTTCTCCAAGA,
(10)CXCL2 For-TTGCCAGCTCTCCTCCTC Rev-CTCAGCAGGCGGTTCGAG, (11)PCSK2 For-TGCAGATTTAGCATCAAGCA Rev-CGGAGAGAAAGAGCGAGTG,
(12)HOXD12 For-ACAGAGCGGGCTATGTGG Rev-GGCGACTGCAGATTCAGA, (13)KCNJ3 For-GTCCCAGGGGAGAAGGAG Rev-GGTAGCTGCGGTCTCTGC,
(14)OR9Q1 For-ACCTTGGTGACCGAGTTCC Rev-GTGCCCATTCAGGATATTCA

Composite Profiles

Composite profiles of enrichments are described previously (Pokholok et al., 2005). Briefly, selected genes are aligned with each other according to the position of their transcription start sites. Aligned probes were then assigned to 100 bp segment bins, and an average of the corresponding enrichment ratio was calculated. Assignment of genes into either absent or present based on affymetrix data, detected or not detected by MPSS analysis, or various categories of relative expression are summarized in Table S3.

While composite profiles provide an easy visualization of large amounts of genome-wide data, subtle, but statistically significant changes can be hidden, particularly with low ratios. Such changes would be expected to alter the distribution of ratios in different portions of the gene and can be detected using a two-sample Wilcoxson rank sum test (Bauer, 1972; Hollander and Wolfe, 1973). We used this test to examine the oligos upstream and downstream of the transcription start to look for differences in H3K36me3 enrichment on genes that are most likely to be silent in human embryonic stem cells. Using oligos between 1.5kb and 2.5kb from the transcription start site of genes that are absent by Affymetrix, not detected in MPSS samples and in the bottom 10% of genes based on relative expression, we found no statistical difference between the distribution of enrichment ratios upstream of the gene and the distribution downstream of the start site (data not shown). A similar analysis using the whole list of absent genes did detect a significant change but we cannot exclude that this difference may be due to a number of incorrect calls based on the expression analysis (see Figure 4).

Sample Preparation and Analysis of Differentiated Cell Types

Hepatocytes were obtained directly from perfused human liver at the University of Pittsburgh through the Liver Tissue Procurement and Distribution System (Dr. Stephen Strom). REH cells were cultured in RPMI 1640 + 10% FBS. Cells were crosslinked as above.

Supplemental Results and Discussion

Effect of Histone Methylation Longevity on Detection Levels

Enrichment signals for RNA polymerase II reflect the relative level of polymerase occupancy at the time of the ChIP assay. This is also true for histone methylation, but because methylated histones may remain at promoter sites long after a transcription initiation event has occurred, the enrichment signal for modified histones will likely be greater than that for RNA polymerase, assuming that the antibodies have similar affinities for these antigens. The longevity of H3K4me3 is on the order of hours (Waterborg, 1993; Ng et al., 2003); therefore, even a brief association of Pol II with a promoter during transcription initiation could lead to a long-lived signal for histone methylation. It is reasonable to hypothesize then, that even where levels of Pol II are below the threshold of detection, the enrichment of H3K4me3 at start sites serves as an indicator of a history of transcription initiation. This suggests that H3K4me3 may be a good proxy for transcription initiation sites genome-wide, and that only a small number of cell types would need to be examined to identify the vast majority of promoters in the genome.

Tissue Specificity of H3K4 Methylation and Association of H3K4me3 with CpG Islands

The majority of all genes, both active and inactive, are H3K4 methylated in each of three cell types (hES, hepatocyte, and B cells). However, a subclass of genes show differences in H3K4me3 state between these cell types. Overall, 75% of the transcription start sites analyzed do not show differences in H3K4me3 occupancy, but the remainder demonstrate some variability. Gene ontology was used to look for any categories significantly enriched in the remaining 25% of genes that show variable expression depending on the specific methylation pattern (http://gostat.wehi.edu.au/cgi-bin/goStat.pl). Genes that were methylated in all cell types were dramatically enriched for all terms associated with metabolism, which can be explained by the large number of housekeeping genes included.

Genes specifically methylated in ES cells included many associated with neural functions including transmission of nerve pulses, synaptic transmission, and ion channel activity (all around 10^{-20} – 10^{-15}) as well as some genes involved in morphogenesis, cell differentiation, and organ development ($4x10^{-9}$) Similarly, genes methylated specifically in B cells and liver samples appear to have cell-type-specific functions. These include oxidoreductase genes in Liver and genes associated with immune defense response in B cells.

The presence of CpG islands within H3K4me3 bound regions correlates differently between the three cell types. CpG islands are frequently associated with transcription start sites and have been correlated with genes involved in basic cellular function as well as early development (Larsen et al., 1992; Ponger et al., 2001; Robinson et al., 2004). Consistent with these results, over 3/4 of the promoters that are constitutively methylated at H3K4 and 1/2 the promoters that are specific to hES cells contain CpG islands. In fact, nearly all the transcription start sites annotated to be at CpG islands are methylated in hES cells (though not all CpG islands are annotated as promoters and only ~2/3 are methylated genome-wide). By comparison, less then 10% of the promoters that are H3K4 trimethylated only in liver or only in B cells contain CpG islands. These results suggest that promoters constitutively H3K4 methylated are much more likely to contain CpG islands, whereas promoters that are differentially H3K4 methylated between different cell types are likely to be distal from CpG islands.

H3K4me3 at Inactive Genes Is Not Limited to Rapidly Inducible Genes

While we have proposed that transcription initiation is widespread at most genes where transcript is not detectable, others who have observed H3K4me3 enrichment at inactive genes have proposed that these genes represent targets for rapid transcription following stimulus (Fernandez et al., 2003; Roh et al., 2006). While we also find that H3K4me3 is enriched at genes that would fit this category, we find numerous examples of H3K4me3 binding at tissue specific targets that are highly unrelated to the examined cell type and unlikely to be rapidly induced. Some examples include enrichment at mitochondrial uncoupling proteins that produce heat in muscle cells (UCP1 and UCP3, Fig2, Table S4), egg receptors for sperm proteins (eg ZP3 and PKDREJ), the blood typing gene ABO, and numerous, non-clustered serotonin receptors (eg. HTR4, HTR7). All of these examples are enriched for H3K4 methylation in all three cell types examined, even though these tissues are unlikely to ever express or encounter these proteins in vivo.

Using GO ontology of all transcriptionally sampled but unexpressed genes (based on MPSS data in hES cells), we find that the most dominant pattern is an anti-enrichment of genes that are in clusters (e.g. olfactory receptors, $p<10^{-22}$). Because these clustered genes represent up to half of all the genes that are not methylated, we anticipate that functional groups that are not found in these clusters, particularly those associated with CpG islands, have a natural tendency to be overrepresented in the bound population.

H3K4 Methylation and Polycomb Group Protein Binding

The inability of many active promoters to produce detectable amounts of mRNAs could be explained by the simultaneous action of transcriptional repressors. One possible mechanism for repression at H3K4me3 enriched genes proposed was the creation of bivalent domains that contain both H3K4me3 and also trimethylation at H3K27 (Bernstein et al., 2006). H3K27me3 is a histone mark created by Polycomb group (PcG) proteins and is associated with silent chromatin. We are able to confirm the presence of PcG proteins at a subset of H3K4me3 enriched genes. Comparing H3K4me3 data to binding of the PcG protein SUZ12 (Lee et al., 2006) shows that over 90% of Polycomb bound domains are also enriched for H3K4me3, but we find that the fraction of genes bound by SUZ12 is too small to account for the majority of H3K4me3 at silent genes. Of 4640 H3K4me3 enriched genes defined as absent based on MPSS data, only 22% are also bound by SUZ12. Thus, incomplete transcription during global transcriptional sampling is likely the function of many modes of repression, one of which is Polycomb mediated silencing.

Most Genes Measured as Absent/Not Detected by Array or MPSS Data Do Not Produce Full-Length Transcripts

Our data suggest that the majority of all genes experience transcription initiation, but only a subset produce detectable, full-length transcript. An alternative interpretation could be that full-length transcripts are produced at most genes, but are simply undetectable by array based methods. We considered the possibility that hES expression data gathered from massively parallel signature sequencing (MPSS) and Affymetrix gene expression arrays (Described above, Table S3) might not accurately determine the absence of full-length transcripts. This could be due to sensitivity issues inherent in the MPSS and Affymetrix methods or differences in cell culture conditions. To verify that genes measured as “absent/not detected” by Affymetrix/MPSS methods were called correctly (do not produce full-length transcript), we measured a subset of genes by reverse transcription coupled to real-time quantitative PCR.

38 H3K4me3 enriched genes called both “absent” by Affymetrix and “not detected” by MPSS method were randomly selected and tested by real time quantitative RT-PCR analysis for the presence of full-length transcript. Since RT-PCR has higher dynamic range than array

based detection (Rockett and Hellmann, 2004), we would expect the majority of the genes to verify the array based results, but some gene transcripts to be detected by RT-PCR. We selected an arbitrary cutoff of 1 transcript per cell as genes that are detected below that rate may be transcribed either stochastically from the hES cell population, from the <10% of cells that may have begun differentiating (see above) or from genomic DNA contamination. As expected, all genes demonstrated some level of detection, however only ~25% of genes called "absent/not detected" were detectable at an estimated $\geq$1 transcript per cell by real time quantitative RT-PCR analysis (Figure 4, Table S6).

H3K79me2 Is a Histone Modification Associated with Elongation

While most of the histone modifications used in this study have been extensively investigated over the last decade, less is known about H3K79 methylation. H3K79 is a residue on the core of histone H3. Methylation of H3K79 is thought to be associated with gene activation and in yeast is thought to facilitate mating type silencing indirectly by inhibiting SIR2/3/4 binding (Khan and Hampsey, 2002; Schubeler et al., 2004; Morillon et al., 2005; Pokholok et al., 2005). In our studies we find that H3K79me2-modified nucleosomes peak immediately downstream of the transcription start site of active genes, but, unlike H3K4me3 and H3K9,14ac, H3K79me2 is limited to active genes. Over two-thirds of active genes in hES cells have high-confidence enrichment of H3K79me2 as compared to less then 10% of genes for which transcripts have not been detected (Table S2). Importantly this suggests that H3K79me2 may be a better marker for actively transcribed genes then the more standard H3K4me3-modification or RNA polymerase II, both of which are present at a large number of inactive genes.

Supplemental References

Abeyta, M. J., Clark, A. T., Rodriguez, R. T., Bodnar, M. S., Pera, R. A., and Firpo, M. T. (2004). Unique gene expression signatures of independently-derived human embryonic stem cell lines. Hum Mol Genet *13*, 601-608.

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. J Am Stat Soc *67*, 687-690.

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

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G*., et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Brandenberger, R., Khrebtukova, I., Thies, R. S., Miura, T., Jingli, C., Puri, R., Vasicek, T., Lebkowski, J., and Rao, M. (2004). MPSS profiling of human embryonic stem cells. BMC Dev Biol *4*, 10.

Brodsky, A. S., Meyer, C. A., Swinburne, I. A., Hall, G., Keenan, B. J., Liu, X. S., Fox, E. A., and Silver, P. A. (2005). Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. Genome Biol *6*, R64.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Cho, E. J., Kobor, M. S., Kim, M., Greenblatt, J., and Buratowski, S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev *15*, 3319-3329.

Cowan, C. A., Klimanskaya, I., McMahon, J., Atienza, J., Witmyer, J., Zucker, J. P., Wang, S., Morton, C. C., McMahon, A. P., Powers, D., and Melton, D. A. (2004). Derivation of embryonic stem-cell lines from human blastocysts. N Engl J Med *350*, 1353-1356.

Davuluri, R. V., Grosse, I., and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. Nat Genet *29*, 412-417.

Fernandez, P. C., Frank, S. R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. (2003). Genomic targets of the human c-Myc protein. Genes Dev *17*, 1115-1129.

Gerhard, D. S., Wagner, L., Feingold, E. A., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., Good, P.*, et al.* (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res *14*, 2121-2127.

Hollander, M., and Wolfe, D. H. (1973). Nonparametric statistical inference. In (John Wiley and Sons), pp. 27-75.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F.*, et al.* (2005). Ensembl 2005. Nucleic Acids Res *33 Database Issue*, D447-453.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D.*, et al.* (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109-126.

Jones, J. C., Phatnani, H. P., Haystead, T. A., MacDonald, J. A., Alam, S. M., and Greenleaf, A. L. (2004). C-terminal repeat domain kinase I phosphorylates Ser2 and Ser5 of RNA polymerase II C-terminal domain repeats. J Biol Chem *279*, 24957-24964.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Khan, A. U., and Hampsey, M. (2002). Connecting the DOTs: covalent histone modifications and the formation of silent chromatin. Trends Genet *18*, 387-389.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876-880.

Kristjuhan, A., Walker, J., Suka, N., Grunstein, M., Roberts, D., Cairns, B. R., and Svejstrup, J. Q. (2002). Transcriptional inhibition of genes with severe histone h3 hypoacetylation in the coding region. Mol Cell *10*, 925-933.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095-1107.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K.*, et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. Embo J *23*, 4051-4060.

Morillon, A., Karabetsou, N., Nair, A., and Mellor, J. (2005). Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription. Mol Cell *18*, 723-734.

Morillon, A., Karabetsou, N., O'Sullivan, J., Kent, N., Proudfoot, N., and Mellor, J. (2003). Isw1 chromatin remodeling ATPase coordinates transcription elongation and termination by RNA polymerase II. Cell *115*, 425-435.

Ng, H. H., Robert, F., Young, R. A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell *11*, 709-719.

Parra, G., Blanco, E., and Guigo, R. (2000). GeneID in Drosophila. Genome Res *10*, 511-515.

Patturajan, M., Conrad, N. K., Bregman, D. B., and Corden, J. L. (1999). Yeast carboxyl-terminal domain kinase I positively and negatively regulates RNA polymerase II carboxyl-terminal domain phosphorylation. J Biol Chem *274*, 27823-27828.

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E.*, et al.* (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. Cell *122*, 517-527.

Pokholok, D. K., Zeitlinger, J., Hannett, N. M., Reynolds, D. B., and Young, R. A. (2006). Activated signal transduction kinases frequently occupy target genes. Science *313*, 533-536.

Ponger, L., Duret, L., and Mouchiroud, D. (2001). Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res *11*, 1854-1860.
Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33 Database Issue*, D501-504.

Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R. A., and Dynlacht, B. D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev *16*, 245-256.

Robinson, P. N., Bohme, U., Lopez, R., Mundlos, S., and Nurnberg, P. (2004). Gene-Ontology analysis reveals association of tissue-specific 5' CpG-island genes with development and embryogenesis. Hum Mol Genet *13*, 1969-1978.
Rockett, J. C., and Hellmann, G. M. (2004). Confirming microarray data--is it really necessary? Genomics *83*, 541-549.

Roh, T. Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A *103*, 15782-15787.

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C., Schreiber, S. L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Sato, N., Sanjuan, I. M., Heke, M., Uchida, M., Naef, F., and Brivanlou, A. H. (2003). Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev Biol *260*, 404-413.

Schubeler, D., MacAlpine, D. M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D. E., O'Neill, L. P., Turner, B. M., Delrow, J.*, et al.* (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev *18*, 1263-1271.

Solter, D., and Knowles, B. B. (1979). Developmental stage-specific antigens during mouse embryogenesis. Curr Top Dev Biol *13 Pt 1*, 139-165.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G.*, et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.

Waterborg, J. H. (1993). Dynamic methylation of alfalfa histone H3. J Biol Chem *268*, 4918-4921.

Wei, C. L., Miura, T., Robson, P., Lim, S. K., Xu, X. Q., Lee, M. Y., Gupta, S., Stanton, L., Luo, Y., Schmitt, J.*, et al.* (2005). Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem Cells *23*, 166-185.

Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H., and Farnham, P. J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev *16*, 235-244.

Supplemental Tables

All Supplemental Tables can be found in the accompanying Microsoft Excel files.

Table S1. Regions of the Human Genome Enriched for H3K4me3 at High Confidence in ES Cells

Table S2. Genes Enriched with High Confidence

Table S3. Maximum Enrichment Ratios of H3K4me3 in hES Cells, Hepatocytes, and B Cells

Table S4. Comparison of H3K4me3 Enriched Genes between Genome-wide and Promoter Arrays

Table S5. Summary of Expression Data from ES Cells

Table S6. mRNA Levels Measured by RT-PCR in hES Cells

Table S7. Regions of the Human Genome Enriched for H3K4me3 at High Confidence in Liver Samples

Table S8. Regions of the Human Genome Enriched for H3K4me3 at High Confidence in B Cells (REH)

Figure S1. Human H9 ES Cells Cultured on a Low Density of Irradiated Murine Embryonic Fibroblasts

Bright-field image of H9 cell culture.

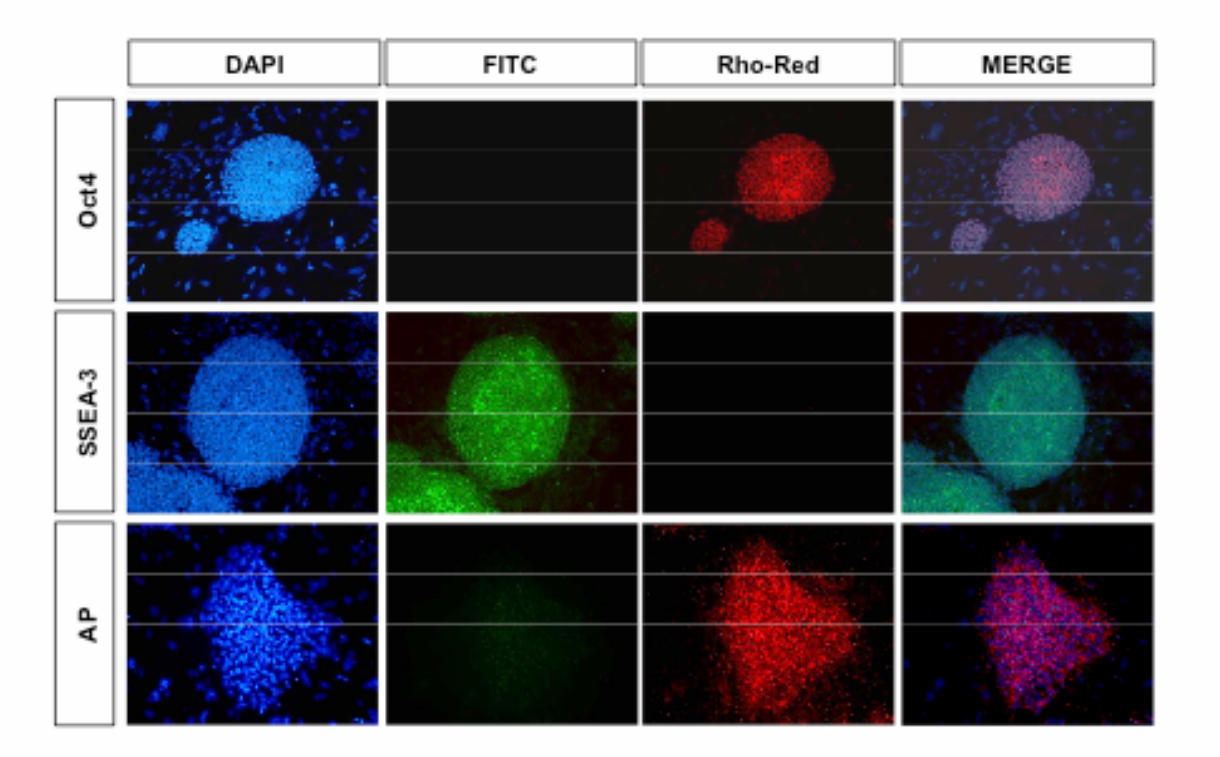

Figure S2. Analysis of Human ES Cells for Markers of Pluripotency

Human embryonic stem cells were analyzed by immunohistochemistry for the characteristic pluripotency markers Oct4 and SSEA-3. For reference, nuclei were stained with DAPI. Our analysis indicated that >90% of the ES cell colonies were positive for Oct4 and SSEA-3. Alkaline phosphatase activity was also strongly detected in human ES cells.

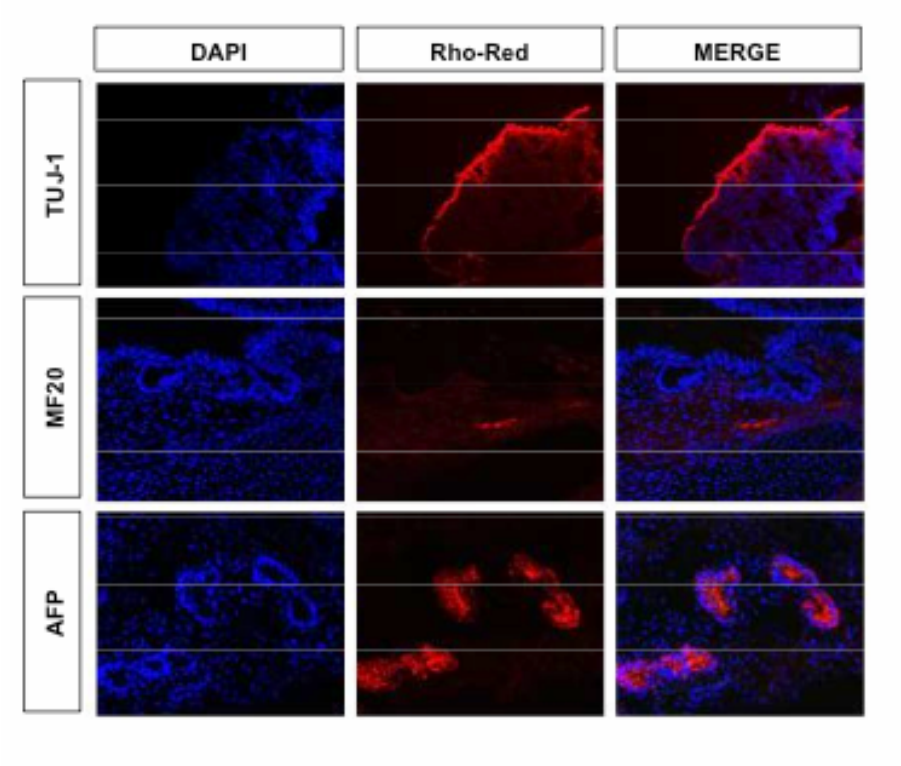

Figure S3. Analysis of Human ES Cells for Differentiation Potential

Teratomas were analyzed for the presence of markers for ectoderm (Tuj1), mesoderm (MF20) and endoderm (AFP). For reference, nuclei are stained with DAPI. Antibody reactivity was detected for derivatives of all three germ layers confirming that the human embryonic stem cells used in our analysis have maintained differentiation potential.

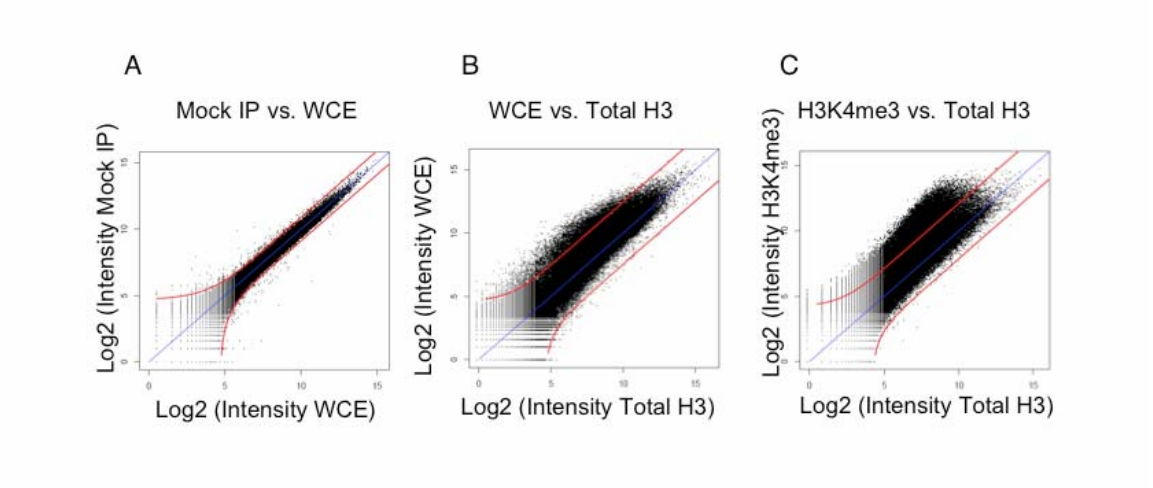

Figure S4. Scatter Plots of Selected ChIP-Chip Experiments

ChIP/chip experiments from human H9 ES cells. A) Log2 intensity for Non-specific IgG enriched DNA compared to Log2 intensity of genomic DNA. 1:1 ratio is indicated by the blue line. Single probe p-value of 10^{-3} is indicated by the red lines. B) Log2 intensity for genomic DNA compared to Log2 intensity of anti-H3 globular enriched DNA. C) Log2 intensity for anti-H3K4me3 enriched DNA compared to Log2 intensity of anti-H3 globular enriched DNA.

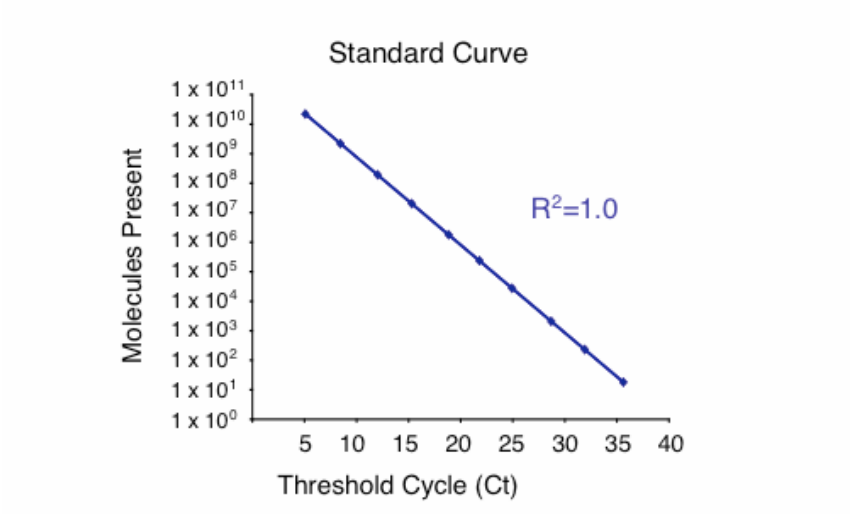

Figure S5. Standard Curve for qPCR

Compiled standard curve of *ANG*, *GPR143* and *KCNJ3* cDNAs using Ct-values derived from quantitative real-time PCR (Applied Biosystems TaqMan assays; 7500 system and analysis).

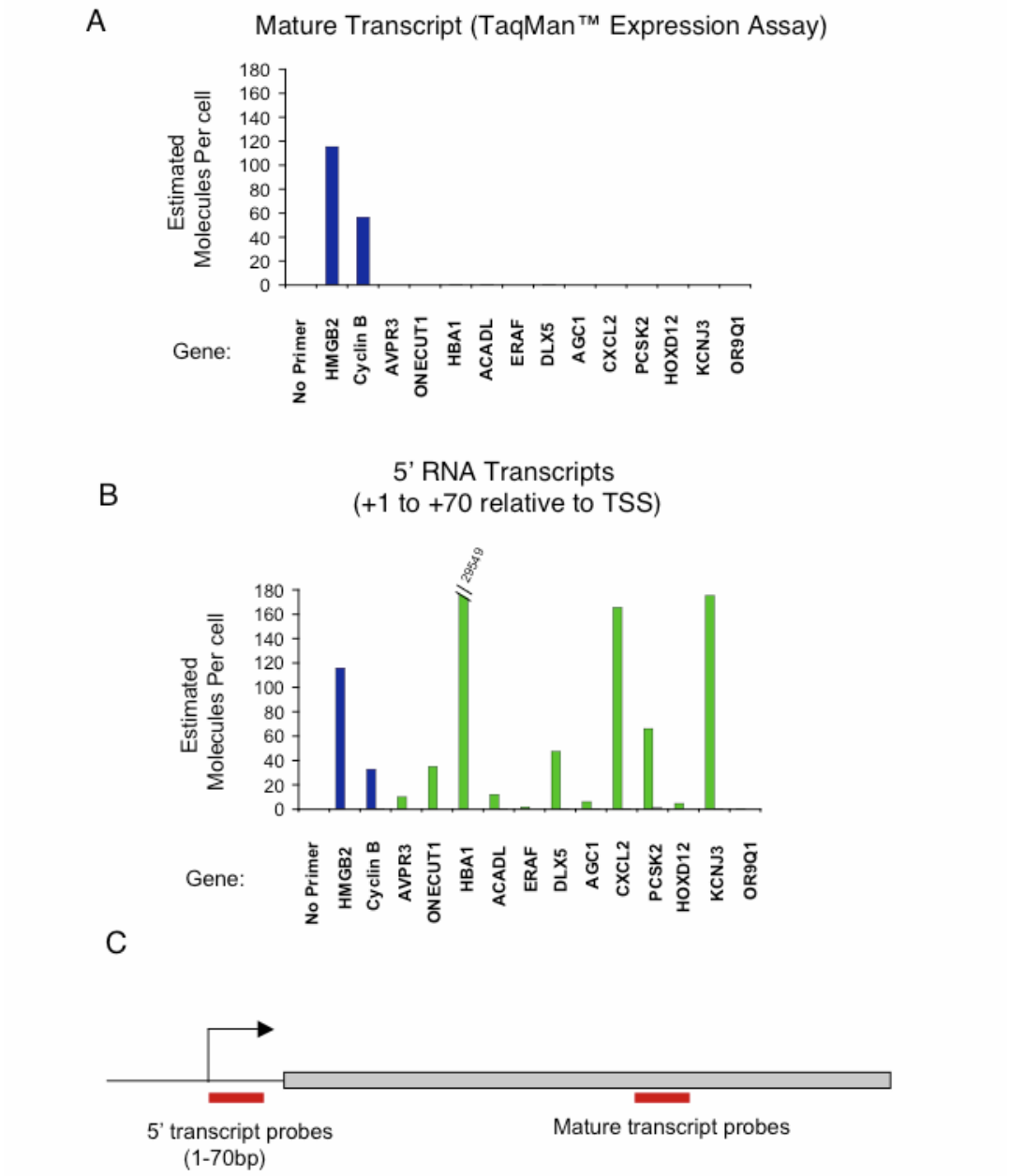

Figure S6. RT-PCR-Based Detection of 5' RNA Transcripts at Inactive Genes

(A) RT-PCR-based detection of full-length mRNAs from total RNA extracted from hES cells using TaqMan probes (Applied Biosystems). Estimated transcripts per cell were obtained by dividing the number of molecules of cDNA, as determined by relative quantitation using the standard curve in (S5) by the number of cells used for the amplification. Positive control probes (active genes as determined by MPSS and Affymetrix data, blue bars) and test sample probes (Inactive genes as determined by MPSS and Affymetrix data, green bars) are as follows: *(1) No probe (2)HMGB2, (4)CCNB1, (5)AVPR3, (6)ONECUT1, (7)HBA1, (8)ACADL, (9)ERAF, (10)DLX5, (11)AGC1, (12)CXCL2, (13)PSK2, (14)HOXD12, (15)KCNJ3, (16)OR9Q1.*

(B) RT-PCR based detection of 5' RNAs from total RNA extracted from hES cells using SYBR green detection (Applied Biosystems). Estimated transcripts per cell were obtained by dividing

the number of molecules of cDNA, as determined by relative quantitation using the standard curve in (S5) by the number of cells used for the amplification. Positive control probes (active genes as determined by MPSS and Affymetrix data, blue bars), control no template (no reverse transcription samples, grey bars which are near zero) and test sample probes (Inactive genes as determined by MPSS and Affymetrix data, green bars) are as follows: *(1) No probe (2)HMGB2, (3)CCNB1, (4)AVPR3, (5)ONECUT1, (6)HBA1, (7)ACADL, (8)ERAF, (9)DLX5, (10)AGC1, (11)CXCL2, (12)PSK2, (13)HOXD12, (14)KCNJ3, (15)OR9Q1.*

(C) Schematic diagram of probe placement for an average gene. 5' RNA was detected within 1-70bp of transcription start site and full-length transcript detection was further downstream.

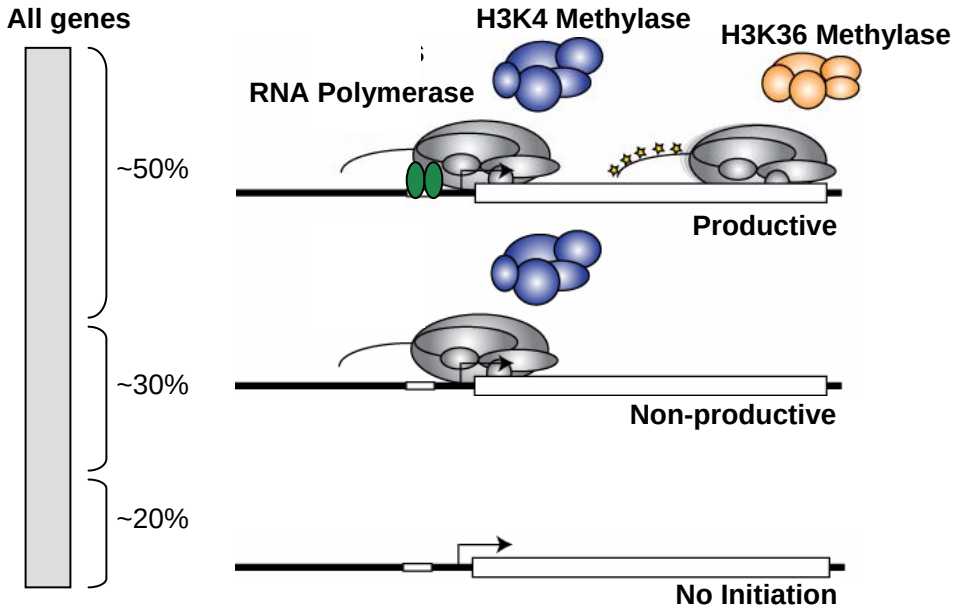

Figure S7. Three Classes of Genes in Human ES Cells

Our results suggest that most protein-coding genes in human cells, including most genes thought to be transcriptionally inactive, experience the hallmarks of transcription initiation. To estimate the fraction of genes in each class, we used a simple voting mechanism to assign every gene as either active or inactive by calling a gene active if one Affymetrix probe for a given gene was detected in at least half of the experiments we analyzed (left, Supplemental text). Using this algorithm, ~50% of genes were considered active, over 90% of which had H3K4me3-modified nucleosomes at one of their promoters in ES cells. Among the remaining half of the genes, 60% had H3K4me3-modified nucleosomes at the transcription start site where regulation of events subsequent to transcription initiation must play important roles in preventing production or accumulation of transcripts. , The final 20% of the total population where the promoters did not contain H3K4me3-modified nucleosomes consists of genes that are excluded from experiencing transcriptional initiation, where mechanisms that prevent transcription initiation must predominate. Schematics of the expected binding for several complexes for three classes are illustrated on the right.

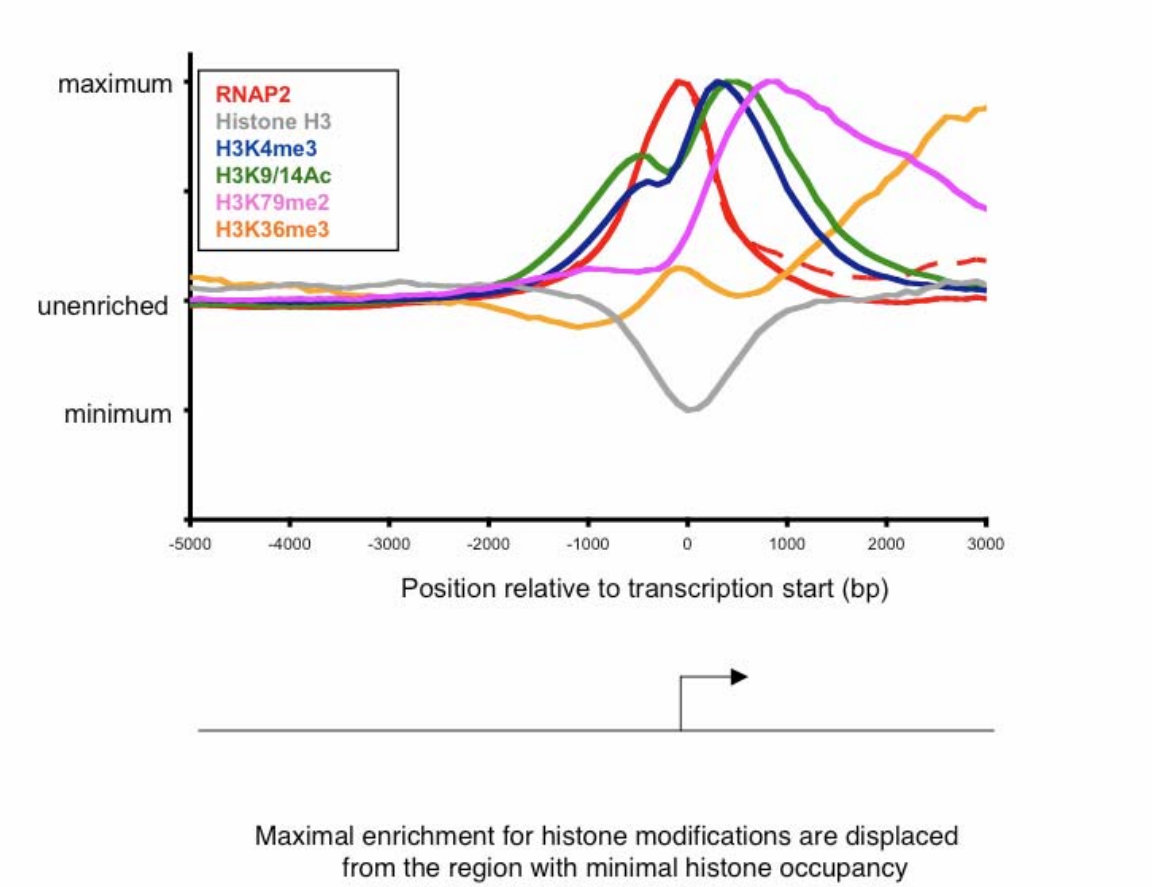

Figure S8. Composite Profile of All Histone Marks at Active Genes

Composite enrichment profiles for genes called present for mRNA transcript using Affymetrix data (Sato et al., 2003; Abeyta et al., 2004). The start and direction of transcription of the average gene is noted by arrow. Experiments included are initiating RNA polymerase II (8WG16, red solid), elongating RNA polymerase II (4H8, red dashed), H3 globular (grey), H3K4me3 (blue), H3K9/14Ac (green), H3K79me2 (pink) and H3K36me3 (orange).

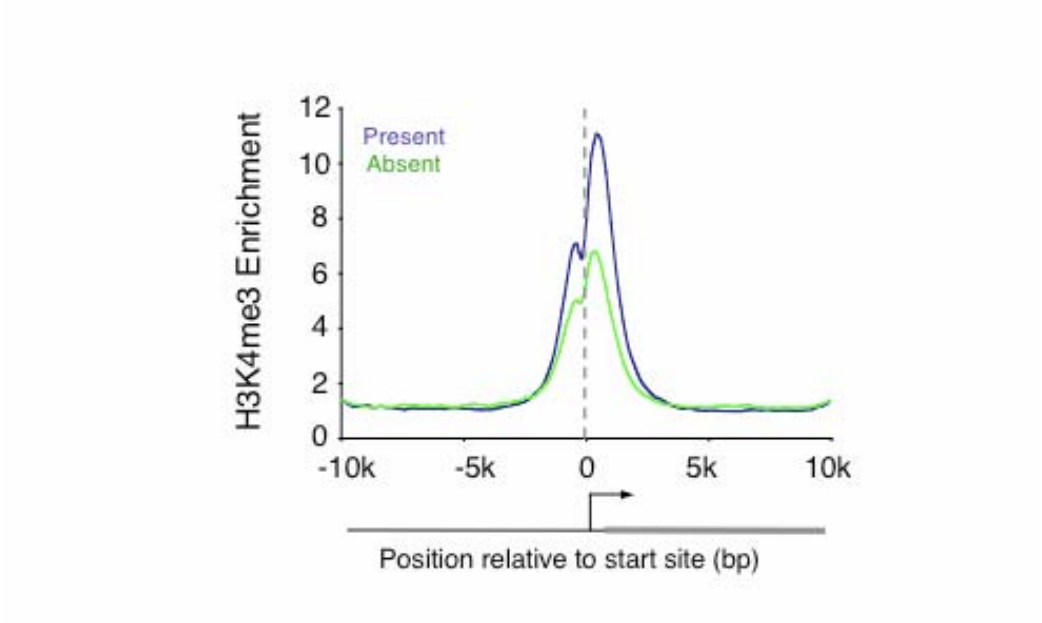

Figure S9. H3K4me3 Enrichment at High-Confidence Targets at Active and Inactive Genes

Composite H3K4me3 enrichment profiles for genes enriched for H3K4me3-modified nucleosomes at high confidence called present (blue) or absent (green) for mRNA transcript using Affymetrix data as in figure 2G (Sato et al., 2003; Abeyta et al., 2004).

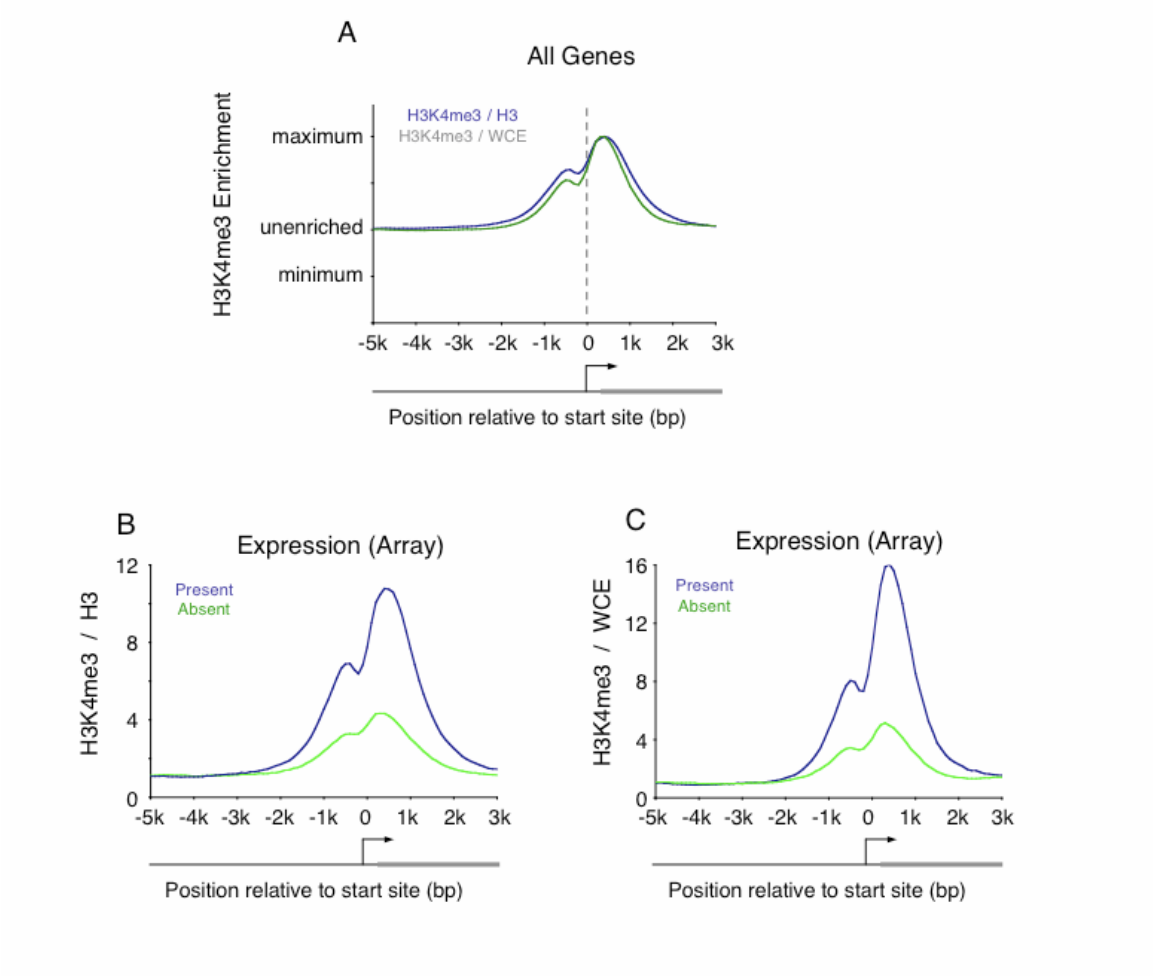

Figure S10. H3K4me3 Enrichment at Inactive Promoters Is Not Due to Depletion of H3

(A) Composite H3K4me3 enrichment profiles for all genes normalized to Histone H3 enrichment (blue) or whole cell extract (green). Experiments are normalized to maximal signal for easier comparison.

(B) Composite H3K4me3 enrichment normalized to Histone H3 enrichment for genes called present (blue) or absent (green) for mRNA transcript using Affymetrix data as in figure 2G (Sato et al., 2003; Abeyta et al., 2004).

(C) Composite H3K4me3 enrichment normalized to whole cell extract profiles for genes called present (blue) or absent (green) for mRNA transcript using Affymetrix data as in figure 2G (Sato et al., 2003; Abeyta et al., 2004).