Supplementary Information

S1 Experimental methods

OVERVIEW OF METHODS

Experimental procedures: HeLa and IMR90 cells were obtained from ATCC and cultured under recommended conditions. Passage 32 H1 cells were grown in mTeSR1[1] medium on Matrigel (BD Biosciences, San Jose, California), for 5 passages. 15 X 10cm^2 dishes were grown using standard mTeSR1 culture conditions and 20 X 10cm^2 dishes were cultured in mTeSR1 supplemented with 200ng/ml BMP4 (RND systems, Minneapolis, MN). 5 days post passage, when cells were approximately 70% confluent, H1 p32 cells grown in unmodified mTeSR1 were crosslinked. To crosslink, 2.5ml of crosslinking buffer (5M NaCl, 0.5M EDTA, 0.5M EGTA, 1M HEPES pH 8, 37% fresh formaldehyde) as added to 10ml culture medium and incubated at 37ºC for 30 minutes, 1.25ml of 2.5M Glycine was added to stop the crosslinking reaction. Cells were removed from culture dish with a cell scraper, and collected by centrifugation for 10 minutes at 2500 rpm at 4C. Cells were washed three times with cold PBS. After the final spin, cells were pelleted and flash frozen using liquid nitrogen. BMP4-treated cells were subjected to the same procedure after 6 days of exposure. GM06990 (#GM06990) B-lymphocyte cells were acquired from Coriell (www.ccr.coriell.org) and grown to a density of 2.5x10^{5}cells/mL in RPMI 1640 medium with 2mM L-glutamine containing 15% fetal bovine serum at 37°C, 5% CO_2. K562 (#CCL-243) cells were acquired from ATCC (www.atcc.org) and grown to a density of 2.5x10^{5}cells/mL in Iscove’s modified Dulbecco’s medium with 4 mM L-glutamine containing 1.5 g/L sodium bicarbonate, and 10% fetal bovine serum at 37°C, 5% CO_2. Chromatin preparation, ChIP, DNA purification, and LM-PCR were performed as previously described[2-4], using commercially available antibodies (α-H3K27ac, Abcam ab4729; α-H3K4me1, Abcam ab8895; α-H3K4me3, Upstate 07-473; α-p300, Santa Cruz sc-585; α-MED1, Santa Cruz sc-5334; α-STAT1, Santa Cruz sc-345). CTCF ChIP was performed with a previously described antibody[3]. ChIP samples were hybridized to the NimbleGen genome-wide tiling microarray set (NimbleGen Systems, Inc.) as previously described[2,3] and to custom condensed enhancer microarrays (NimbleGen Systems, Inc.) using standard methods. The condensed enhancer microarrays consisted of tiled 10 kb windows around each of 38716 primary predicted enhancers and standard controls. DNase-chip was performed and the data analyzed as previously described[5]. Cloning and reporter assays were performed as previously described[4] and a fragment was designated as active if its relative luciferase value was greater than 2.33 standard deviations ($p = 0.01$) above the median random activity.

ChIP-chip data analysis and chromatin signature-based predictions: Data were analyzed using standard methods, and ChIP-chip targets for CTCF, p300, MED1, and STAT1 were selected with the Mpeak program. Enhancers were predicted as previously described[6], with slight modifications to account for probe spacing on genome-wide array platforms. Genome-wide enhancer predictions in HeLa cells were considered to be verified if their averaged H3K4me1 and H3K4me3 enrichment profiles on the condensed enhancer microarrays were sufficiently correlated to known enhancer chromatin signatures, resulting in verification of 36589 enhancers out of 38716 primary predictions (94.5% verified). In K562 cells, we performed ChIP-chip on Nimblegen HD2 genome-wide tiling arrays (12 array set, hg18). We normalized the raw data from each array using MA2C[7], and mapped the normalized data to hg17 coordinates using the UCSC liftOver tool. We predicted K562 enhancers using the H3K4me1 and H3K4me3 profiles as in HeLa, with the following changes: we could not simply map the HeLa training set used in Heintzman et al to the K562 dataset, as the cell types are different. Instead, we used the unaltered training set from Heintzman et al. ROC analysis (data not shown) indicates that a correlation cutoff of 10% and an intensity cutoff of 1 X 10^{-12} yields the best overlap with K562 predictions in the ENCODE regions. K-means clustering, intersection analysis, evolutionary conservation analysis, and other computational comparisons (to UCSC Known Genes, ChIP-chip target lists, etc.) of the prediction sets were performed as previously described[6]. Specifically, to assess the overlap of predicted enhancers with genome-wide transcription factor binding site (TFBS) data sets, we counted the number of experimentally determined TFBS within 2.5 kb of the enhancers. To determine the significance of this overlap, we compared this statistic to the distribution of statistics for 100 random sets of putative enhancers, which was approximated by a normal. Each random set had the same number of elements as the putative enhancer set. Our enhancer predictions were limited to regions on the ChIP-chip array. Similarly, each random enhancer was placed uniformly at random in a sample space consisting of well-represented regions on the ChIP-chip microarray. The chromosomal distribution of each of the sets was kept constant. This careful placement of random sites ensures we do not artificially inflate the significance of the overlaps.

Gene expression and entropy analysis: Gene expression in the various cell lines was analyzed using HGU133 Plus 2.0 microarrays (Affymetrix) as described[6]. Specificity of expression was determined using a function of Shannon entropy as described[8] and the top, middle, and bottom 1000 genes from this analysis were designated as HeLa-specific expressed, non-specific expressed, and HeLa-specific repressed genes, respectively (Figure S5), for evaluation of enhancer enrichment in the insulator-defined domains containing the promoters for these classes of genes (as in Figure 3A), where insulators were defined by CTCF binding sites as described below. When counting enhancers around these promoters, we included all enhancers within 200 kb of a promoter as long as they were still within the same insulator-defined domain as described above. Random distributions were generated by averaging the enrichment profiles around promoters of 100 iterations of randomly selected enhancer sets of 36589 elements. To assess enhancer and gene expression specificity between HeLa and K562 cells (as in Figure 3E), we use the MAS5 algorithm from the Bioconductor R package to generate gene expression Present/Absent calls from each cell type. Since we have two biological replicates of K562 expression, to merge calls for these replicates, probes called differently in the two replicates are labeled as Marginal. To eliminate biases from genes not expressed in either HeLa or K562, we only consider a probe if it is called Present in either HeLa or K562. We map Affymetrix probes to gene identifiers using the knownToU133Plus2 table from the UCSC Genome Browser, and then map the identifiers to genomic coordinates (hg17, NCBI build 35) using the knownGene table from the UCSC Genome Browser. To reduce redundancy, we keep only the first gene when multiple Affymetrix probes map to the same annotated gene. The result from this filtering and mapping is a set of 11783 genes. For each such gene, we count the number of enhancers predicted in each cell type within that gene's CTCF-domain. We sort the genes by differential (HeLa – K562) gene expression (as defined by the RMA algorithm from the Bioconductor R package) and use a sliding window of 1000 genes to generate a profile of the average number of enhancers for each cell type as a function of average differential gene expression. This gives two profiles: one using HeLa enhancers and one using K562 enhancers. To normalize, we repeat this analysis for each cell type using 100 sets of random enhancers (placed uniformly at random on the tiling microarrays), giving 100 random enhancer-expression profiles. We then define the enhancer enrichment profile as the ratio between the number of enhancers in the observed profile and the expected number of enhancers in the averaged random profile.

Motif analysis: Enhancer regions were defined as 2 kb windows centered on each prediction, and promoter regions were defined as 1 kb windows upstream from annotated TSS. Promoters regions were excluded from enhancer regions; repeats, exons and transposons were excluded from both. Motif conservation in each region was evaluated relative to the genomes of opossum, tenrec, elephant, armadillo, cow, dog, rabbit, rat and mouse, extracted from UCSC Genome Browser and used with permission. The mammalian tree, along with branch lengths, was computed using DNAML (PHYLIP package)[9] with the F84 nucleotide model of evolution in ~500kb of randomly selected exon sequence. Known and novel motifs were discovered as previously described[10], with the primary difference that instances were not required to have perfect conservation and were considered conserved if they were found across a number of species spanning at least 50% of the total branch length of the mammalian tree (Branch-Length-Score > 50%)[11,12]. We ranked motifs based on their over-conservation, measured as the probability of observing a substantially increased number of conserved motif instances compared to that expected for motifs of identical composition, and selected all motifs with $P < 1 \times 10^{-3}$. We evaluated a motif's enrichment as its over-abundance, or the hypergeometric probability of observing a substantially increased number of occurrences in the intergenic and intronic regions of the human genome (regardless of evolutionary conservation) compared to motifs of identical composition, with a cutoff of $P < 1 \times 10^{-3}$.

Supplementary data for the microarray experiments has been formatted for viewing in the UCSC genome browser via http://bioinformatics-renlab.ucsd.edu/enhancer

S1.1 ChIP-chip

We performed ChIP-chip analysis[13] to determine the chromatin modification patterns along 44 human loci selected by the ENCODE Consortium as common targets for genomic analysis[6], totaling 30 Mbp or 1% of the human genome. We investigated the patterns of six specific histone modifications: acetylated histone H3 lysine 9, 18 and 27 (H3K9Ac, H3K18Ac and H3K27Ac), and mono-, di- and tri-methylated histone H3 lysine 4 (H3K4Me1, H3K4Me2, and H3K4Me3), as well as the insulator-binding protein, CTCF. ChIP samples were amplified, labeled, and hybridized to tiling oligonucleotide microarrays and data were analyzed as previously described[13], generating high-resolution maps of histone modifications and CTCF binding in 5 cell lines: cervical carcinoma HeLa, immortalized lymphoblast GM06690 (GM), leukemia K562, embryonic stem cells (ES), and BMP4-induced ES cells (dES).

ChIP-chip procedure and antibodies against p300, TAF1, histone H3, H3K4Me1, H3K4Me2, H3K4Me3, and CTCF were previously described[2,3,13]. Additional antibodies are commercially available (α-H3K9Ac Abcam ab4441; α-H3K18Ac Abcam ab1191; and α-H3K27Ac Abcam ab4729). All ChIP-chip experiments were completed in triplicate, except for with normal and BMP4-treated ES cells (Figure S8). All ChIP-DNA samples were hybridized to NimbleGen ENCODE HG17 microarrays (NimbleGen Systems). DNA was labeled according to NimbleGen Systems' protocol. Samples were hybridized at 42°C for 16 hours on a MAUI 12-bay hybridization station (BioMicro Systems). Microarrays were washed, scanned and stripped for re-use following protocols from NimbleGen Systems. Gene expression data for HeLa, K562, and GM cells were obtained using HU133 Plus 2.0 microarrays (Affymetrix).

We also performed ChIP-chip of histone modifications H3K4Me1 and H3K4Me3 in HeLa cells on Nimblegen genome-wide tiling arrays (38 array set, hg17). We normalized the raw data from each array using both the median and loess algorithms from the Bioconductor R package (treating each probe equally). For each array, we chose the normalization method that gave the most balanced distribution of random probes about a log ratio of 0. We performed ChIP-chip of histone modifications H3K4Me1, H3K4Me3, and H3K27Ac in K562 cells on Nimblegen HD2 genome-wide tiling arrays (12 array set, hg18). We normalized the raw data from each array using MA2C[7], and mapped the normalized data to hg17 coordinates using the UCSC Genome Browser liftOver tool.

S1.2 Gene Expression Analysis for ES and dES cells

The Human Whole Genome Expression arrays containing ~385,000 60-mer probes was manufactured by NimbleGen Systems (http://www.nimblegen.com). This array design tiles transcripts from approximately 36,000 human locus identifiers for the hg17 (UCSC) assembly with typically 10 or 11 probes per transcript. For gene expression analysis, we isolated the total RNA from H1 ES cells or BMP4-treated cells using Trizol (Invitrogen, Carlsbad, CA) according to the manufacturer’s recommendations. Total RNA was enriched for the polyA fraction using Oligotex mRNA Mini Kit (Qiagen). Enriched mRNA (250 ng) was primed using random hexamers and reverse transcribed using Superscript III (Invitrogen) in the presence of 5-(3-aminoallyl)-dUTP (Ambion). The purified product was coupled to Cy5-NHS ester (Amersham). Similarly, sonicated genomic DNA (2 μg) was primed with random octamers and labeled using Klenow in the presence of 5-(3-aminoallyl)-dUTP. The resulting product was coupled to Cy3-NHS ester (Amersham). Cy3-labeled genomic DNA (4.5 μg) was used as a reference and added along with the Cy5-labeled mRNA sample (2 μg) onto each array. Hybridizations were performed in 3.6X SSC buffer with 35% formamide and 0.07% SDS at 42°C overnight. Arrays were then washed, dried, and scanned using a GenePix 4000B scanner. We set the expression level of genes in undifferentiated cells as 1 and calculated the relative fold change of individual genes in the dES cells.

S2 Data analysis

S2.1 Expression array analysis

We use the GCRMA package[14] to normalize Affymetrix mRNA expression arrays for HeLa, GM, and K562 cell types. For every pair of these cell types, we also use GCRMA to find differentially expressed and repressed genes using a p-value cutoff of 0.01 in conjunction with a fold change cutoff of 2.0. The expression data for ES and dES cell types was done using the Nimblegen platform, and thus are not directly comparable to the Affymetrix expression data. As such, we can only use this expression data to compare ES and dES cell types. As a conservative measure of differential expression, we use a fold-change cutoff of 2.

S2.2 Gene Expression Data Analysis for ES and dES cells

Gene expression raw data were extracted using NimbleScan software v2.1. Considering that the signal distribution of the RNA sample is distinct from that of the gDNA sample, the signal intensities from RNA channels in all eight arrays were normalized with the Robust Multiple-chip Analysis (RMA) algorithm[14]. Separately, the same normalization procedure was performed on those from the gDNA samples. For a given gene, the median-adjusted ratio between its normalized intensity from the RNA channel and that from the gDNA channel was then calculated as follows:

Ratio = intensity from RNA channel/(intensity from gDNA channel + median intensity of all genes from the gDNA channel).

We have found that this median-adjusted ratio gives the most consistent results when compared to other published human ES cell expression data, such as SAGE library information available from the Cancer Genome Anatomy Project (CGAP). Consequently, we used this median-adjusted ratio as the measurement for the gene expression level

S2.3 Expanded maps of histone modifications

Previously, we demonstrated enhancers could be determined by distinct chromatin signatures of H3K4Me1 and H3K4Me3 at these functional elements [13]. Focusing on HeLa cells, we found that three additional histone modification marks, namely H3K9Ac, H3K18Ac and H3K27Ac are also part of the chromatin patterns at promoters and enhancers. All three acetylation marks localize to active transcription start sites (TSSs), and remain absent, as do other chromatin modifications, at inactive promoters (Figure S9A). These results agree with individual promoter studies observing acetylation or hyper-acetylation at active promoters[15,16], as well as with large-scale histone modification studies in yeast[17,18]. Distal p300 binding sites show clear enrichment of H3K18Ac and H3K27Ac, while H3K9Ac is much reduced (Figure S9B). These results suggest that H3K9Ac is preferentially associated with active promoters, while H3K18Ac and H3K27Ac are associated with both promoters and enhancers.

S2.4 Identification of CTCF and p300 binding sites

The Mpeak program can reliably detect binding sites of transcription factors, and has worked well in previous studies to identify TAF1, CTCF, and p300 binding sites[2,3,13,19]. We use the Mpeak program to determine binding sites of CTCF[3] and p300[13] peaks. Specifically, we call a CTCF peak if there is a stretch of 4 probes separated by at most 300 bp that are at least 2.5 standard deviations above the mean. For p300, we use a simple FDR cutoff of 0.0001 to define peaks as in Heintzman et al[13]. We used different parameters for consistency with previous publications, but swapping these parameters does not vary the results significantly.

S2.5 Most human promoters are universally associated with a set of active chromatin marks in different cell types

Modulation of chromatin state is a key component of tissue-specific gene expression programs[17,20]. Given the diversity of these five cell lines and the critical role of promoters in regulating gene expression, we hypothesized that the chromatin modifications at promoters would uniquely define each cell type, but we actually observed the opposite. At promoters of 414 genes in the ENCODE regions, we found that the chromatin signatures at promoters are remarkably similar across all cell types (Figure 1A). To quantify this, we defined a cell type's enrichment profile as the sum of the log ratio enrichment values of H3K4Me1, H3K4Me3, and H3K27Ac for each promoter. We then calculated the Pearson correlation coefficient between enrichment profiles from different cell types (Figure S1A). The enrichment profiles are highly correlated between all pairs of cell types, with an average correlation coefficient of 0.71. This observation also holds at the larger set of Gencode promoters (Figure S2). The generally invariant nature of the chromatin marks at promoters suggest that epigenetic features at this class of regulatory element are not the dominant drivers of cell type-specific gene expression patterns.

S2.6 CTCF binding in the genome is generally cell-type invariant

Insulator elements play key roles in restricting enhancers from activating inappropriate promoters, thereby defining the boundaries of gene regulatory domains[21]. Nearly all insulators that have been experimentally defined in the mammalian genome require the insulator binding protein CTCF to function[22]. Our previous genome-wide location analysis of the insulator binding protein CTCF in human fibroblasts indicated that predicted insulators (those sites in the genome bound by CTCF) are closely correlated with the distribution of genes, and are highly conserved throughout evolution, consistent with their key role in transcription regulation[3]. Intriguingly, the overlap of predicted insulators in two cell lines in that study (IMR90 lung fibroblast and U937 hematopoietic progenitor cells) was a remarkable 67%, suggesting cell-type invariance. To further investigate this possibility, we used ChIP-chip to identify CTCF binding sites in the ENCODE regions in each of the five cell types. On average, 517 predicted insulators were recovered in each cell type, with a remarkable average of 82.8% shared between pairs of cell types (Table S1). Indeed, the CTCF enrichment profiles at 729 non-redundant CTCF binding sites are nearly identical across all five cell types studied here and IMR90 cells (Figure S1E), and the average Pearson correlation coefficient between all pairs of profiles is 0.72 (Figure S1B), comparable to the value observed at promoters. The consistency of CTCF binding appears to extend to the entire genome (Figure S6). These results support insulators as being largely cell-type invariant, to a greater degree than previously appreciated. Additionally, none of the histone modifications that we examined were consistently present at predicted insulators (Figure S11) – further experiments are necessary to determine a chromatin signature for insulators, if one exists.

S2.7 p300 binding sites are cell-type specific

As an alternative method to predicting enhancers based on chromatin modifications, we use the stringent criteria of defining enhancers to be binding sites of p300, a histone acetyltransferase and co-activator protein. While the presence of p300 is sufficient to indicate an enhancer, p300 is not necessarily found at all enhancers[13]. Using the Mpeak program[19], we identified a total of 411 TSS-distal p300 binding sites in HeLa, GM, and K562 cell lines. We observe that, unlike CTCF and chromatin modifications at promoters, the localization of p300 binding sites appears unique to each cell-type in the three cell types where p300 ChIP-chip analysis was performed (Figure 1F). The cell-type specificity of p300 binding sites is supported by the extremely low correlations observed: the average pair wise Pearson correlation coefficient at p300 binding sites is -0.11 (Figure S1C), compared to the much higher correlations 0.71 and 0.72 observed at promoters and insulators, respectively. More strikingly, p300 binding sites are largely cell-type specific: of the 411 distal peaks recovered from the three cell types, the vast majority (378, 92.9%) are unique to a single cell type, 29 (7.1%) are shared among exactly two cell types, and 4 (1.0%) are common among all three cell types. These results support the findings for enhancers defined by chromatin-based signatures.

S2.8 Enhancer prediction method

The procedure used to predict enhancers follows closely to that in Heintzman et al[13]. We first bin the tiling ChIP-chip data into 100 bp bins, averaging multiple probes that fall into the same bin.

Using a sliding window on H3K4Me1 and H3K4Me3, we scan for chromatin signatures resembling a training set of enhancer patterns defined by the p300 binding sites in HeLa cells, keeping only those windows that correlate most with the training sets and that have significant enrichment of chromatin modifications. We use a discriminative filter to keep only those predictions that correlate with an averaged enhancer training set more than the promoter training set. Finally, we apply a descriptive filter, keeping only those remaining predictions having a correlation of at least 0.5 with an averaged training set.

In both ENCODE and genome-wide predictions of this study, we made predictions of active promoters and enhancers as in Heintzman et al[13], with the following modifications:

- Repetitive regions of the genome are not covered by the probes on tiling arrays, contributing to gaps in coverage. In Heintzman et al, we interpolated through all gaps. But this can lead to false positive predictions or biasing of the underlying background distributions when there are many gaps. To remove these concerns, here we interpolate only through gaps smaller than 1000 bp.
- In the prediction algorithm, we slide a 10 kb window across the tiled regions and compute 2 statistics for each window: the correlation with a training set and the sum of the absolute values of intensities of the middle 2 kb region of the window. The correlation part has remained unchanged in this study. In Heintzman et al, the intensity statistic appeared normally distributed, and as such we approximated it with a Gaussian distribution. In light of the larger datasets in this study, this normal assumption did not appear entirely correct. Here, we change our intensity statistic to the sum of squares of the normalized intensities in the 2 kb region. A normalized intensity is an intensity subtracted from the mean array intensity and

divided by the standard deviation of the array intensity. Since each array is properly normalized to follow a Gaussian distribution, by definition, this statistic follows a Chi-squared distribution with 42 degrees of freedom (for each window, each of the 2 modifications has 21 normalized intensities squared: 10 in each direction and one at 0).

The training set used here contained the same six groups of training sets used in Heintzman et al, with the exception that the HeLa enhancer predictions used data derived from the genome-wide H3K4Me1 and H3K4Me3 arrays.

In the ENCODE regions, as in Heintzman et al, we keep predictions in the top 10% of the intensity distribution and top 1% of the correlation distribution. For the genome-wide enhancer predictions in HeLa, a ROC analysis (data not shown) indicates that a correlation cutoff of 1% and an intensity cutoff of 1% yields the best overlap with our previously published predictions in the ENCODE regions. Similarly, for the genome-wide predictions for the MA2C-normalized K562 data, ROC analysis suggests using a correlation cutoff of 10% and an intensity cutoff of 1e-12.

S2.9 Validation of predicted enhancers in the ENCODE regions

Several lines of evidence suggest that the histone-modification-based predictions of enhancers are truly enhancers. First, we compare the predicted enhancers to DNase I hypersensitive (HS) sites, as hypersensitivity is a hallmark of enhancers. Using a recently published set of HS sites[23] mapped in HeLa, GM, K562, and H9 ES cells, we computed the percentage of predicted enhancers within 2.5 kb of HS sites (Figure S3A-D). For comparison, we also computed the overlap percentage of 100 sets of randomly placed enhancers restricted to regions on the ChIP-chip microarray. We notice that predicted enhancers in HeLa (53.0% overlap, Z-score = 20.4, p = 3.2E-93), GM (38.2% overlap, Z-score = 14.4, p = 5.1E-47), K562 (overlap = 62.6%, Z-score = 22.7, p = 3.9E-114), and ES (59.2% overlap, Z-score = 18.0, p = 1.0E-72) are enriched in HS sites in their respective cell types. Thus, the predicted enhancers are supported by HS data.

Second, enhancers are defined to be regions in the genome bound to transcription factors and co-activators. To verify the predicted enhancers, we compare their overlap with p300 binding sites. For every cell line where we mapped p300 binding, we observe significant enrichment of predicted enhancers at p300 binding sites (HeLa: 86.4% overlap, Z-score =27.7 , p = 2.9E-169; GM: 79.2% overlap, Z-score = 35.7, p = 4.6E-279; K562: 63.6% overlap, Z-score = 23.3, p = 1.7E-120) (Figure S3E-G), again supporting the predicted enhancers as being real. To further validate the predicted enhancers in the ES cell line, we rely on the definition of enhancers as binding sites for transcription factors and compare the predicted enhancers with previously mapped binding sites for the ES-specific transcription factors Oct4, Sox2, and Nanog[24] (Figure S4). Compared to predicted enhancers from other cell types, we notice greater than 2-fold enrichment of the predicted ES enhancers with these ES-specific factors. Although we do not have the corresponding functional data for the dES cell type, several lines of evidence suggest that they are also real. First, like the other cell types, the histone modification patterns at predicted dES enhancers are enriched in H3K4Me1 and H3K27Ac, but lack H3K4Me3. Second, there is a significant enrichment of dES enhancers at HS sites and p300 binding sites from the other cell types, indicating that at least some of these dES enhancers are real.

S2.10 Validation of genome-wide predictions of enhancers

The features of enhancers in the HeLa genome are consistent with what we observed across cell types in the ENCODE regions. Most predicted enhancers (23686, 64.7%, p = 6.6e-208) exhibit strong evolutionary conservation (PhastCons ≥ 0.8, see Methods). The genomic distribution of the predicted enhancers are distinct from those of promoters: except for a small fraction that overlap with Known Gene 5'-ends, CAGE tags, or CpG islands, the predicted enhancers are distal to promoters, with predominantly intronic (37.9%) or intergenic (56.3%) localization (Figure 2C). Most predicted enhancers (61.4%) are marked by moderate or high levels of acetylation of H3K27 (Figure 2A). The co-activator p300 and Mediator component MED1, known to bind enhancers, are found at 10741 (29.4%) and 5764 (15.8%) enhancer predictions, respectively (see Methods). Additionally, 19776 (54.1%) of the predicted enhancers exhibit significant DNaseI hypersensitivity. Collectively, we found that 23722 (64.8%) predicted enhancers are supported by some combination of DHS and/or binding of p300 and/or MED1 (Figure 2D). Further, the predicted enhancers seem to be distinct from other distal regulatory elements. Only 2666 (8.0%) enhancers are found near a collection of 23267 TSS-distal CTCF sites called in HeLa, IMR90, and CD4 T cells[3,25,26] (1.53-fold enrichment, p = 7.81e-120). Comparison to a genome-wide binding profile of the repressor NRSF/REST[27] (which binds mainly transcriptional silencer elements) revealed that only 39 (0.11%) predicted enhancers overlap with distal NRSF/REST binding sites, significantly lower than that expected at random (3.23-fold depletion, p =3.21e-12). These findings indicate that our map of predicted enhancers is strongly enriched for true enhancer elements.

S2.11 Identification of novel sequence motifs in enhancers

We examined the predicted HeLa genome-wide enhancers for the presence of DNA sequence elements that may guide the establishment and maintenance of chromatin structure or the recruitment of regulatory factors. We reasoned that if such functional motifs are abundant within enhancers, they could show increased evolutionary conservation across related mammals. Indeed, we found increased conservation of motif-like sequence patterns in enhancer regions, evaluated using several hundred shuffled TRANSFAC motifs across 10 mammals in a phylogenetic framework that tolerates motif movement, partial motif loss, and sequencing or alignment discrepancies. Enhancers showed conservation for 4.3% of instances (at Branch-Length-Score > 50%), which is substantially greater than for the remaining intergenic regions (2.9%, p < 1e-100), and even promoter regions (3.9%, p = 1e-57). This suggests that these predicted enhancer regions may indeed contain functional regulatory motifs.

Consequently, we asked whether motifs for known transcriptional regulators show increased abundance and conservation in enhancer regions. We tested a list of 123 unique TRANSFAC motifs as reported previously[10] and found that 67 (54%) of these motifs are over-conserved in enhancers, and 39 (32%) are enriched in enhancers (Table S12). This suggests that many known motifs for well-studied transcriptional regulators at promoters are likely to also play roles in enhancers, implying strongly shared regulatory mechanisms between these two classes of elements at the DNA sequence level. Indeed, of the 67 known motifs over-conserved in enhancers and the 65 over-conserved in promoters, 54 (83%) are over-conserved in both. The enriched motifs include known sequence motifs for binding of transcription factors involved in diverse cellular processes.

Additionally, we searched for evidence of unique enhancer-specific sequence motifs that have previously remained elusive due to the lack of genome-wide knowledge of enhancers. We performed de novo motif discovery in enhancer regions using multiple alignments of ten mammalian genomes, revealing 41 enhancer motifs, of which 19 match known transcription factor motifs while 22 are novel (Table S13). These motifs show conservation rates between 7% and 22% in enhancers (median 9.3%), compared to only 1.1% for control shuffled motifs of identical composition. Even without taking conservation into account, 27 (65%) of these motifs show significant enrichment in human enhancers. Further, over 90% of these motifs appear to be unique to enhancers, as only 4 motifs are enriched in promoter regions and 12 are in fact depleted in promoters (Table S13). In contrast, shuffled versions of these motifs show significantly reduced enrichment in enhancer regions (only 12% of shuffled motifs, a 5-fold reduction) and also reduced depletion in promoters (22%, a 2-fold increase). This indicates that although enhancer regions contain many known promoter motifs, they also contain unique regulatory sequences that may be specific to enhancer function.

S2.12 CTCF knockdown analysis

Most of the predicted enhancers (92%) are located greater than 10 kb from the nearest transcription start site (TSS), posing a challenge in assigned enhancers to their appropriate target genes. We partly resolved the enhancer/target gene relationship by using genome-wide location data for the insulator binding protein CTCF[3,25,26]. To determine if CTCF binding sites can be used to define the boundaries of regulatory domains within which enhancers and gene promoters may interact, we examined the effects of the loss of CTCF on global gene expression. A recent study showed that siRNA-mediated CTCF depletion in HeLa cells resulted in upregulation of

1062 genes and downregulation of 1130 genes (at least 50% change)[25]. We hypothesized that upregulation of these genes was caused by increased interactions of their promoters with nearby enhancers that had been blocked by CTCF prior to its depletion (Figure S14, upper panel), in line with the current understanding of CTCF function. If so, we expect to find more predicted enhancers in the vicinity of upregulated genes and fewer enhancers near genes whose expression is unchanged or downregulated. To test this hypothesis, first we identified insulator-delineated domains in the genome, defining our set of insulators as the union of published CTCF binding sites from IMR90, HeLa, and CD4+ T cells[3,25,26], since we observed consistent CTCF enrichment at nearly all putative insulators across cell types in the ENCODE regions and genome-wide (Figure S1E, Figure S14). We then created three sets of genes: the 1000 most up-regulated upon siCTCF treatment, the 1000 most down-regulated, and the median 1000 unchanged genes. Then we counted predicted enhancers within five insulator-delineated domains (between CTCF binding sites) adjacent to subsets of genes that were upregulated, downregulated, or unchanged by depletion of CTCF in HeLa cells. To generate a random distribution, we also repeated this analysis for 100 sets of 1000 random genes. To obtain enhancer enrichment, we divided the observed counts with the averaged random counts. Finally, to assess significance, we assumed the random counts followed a normal distribution. Indeed, we observed on average a 2.2-fold enrichment of enhancers within domains adjacent to upregulated genes compared to downregulated genes (Figure S14), and a 1.4-fold depletion of enhancers in domains adjacent to downregulated genes relative to genes whose expression is unchanged by CTCF depletion (see Supplemental Materials). These results support the use of CTCF binding sites as boundaries of regulatory domains on a global scale.

S2.13 Enhancers are clustered

To obtain a coarse view of the localization pattern of enhancers, we first examined the distribution of distances between adjacent enhancers. We observed that enhancers are more highly clustered than expected at random (Wilcoxon $p < 1E-300$) (Figure 2E), as has also been observed in Drosophila [28]. Similar analysis from multiple cell types in the ENCODE regions supports this conclusion (Wilcoxon $p = 1.1E-27$) (Figure S12A). In comparison, we observed an enrichment of small TSS-TSS distances, indicative of clustering of TSSs (Wilcoxon $p = 0$, Matlab) (Figure S11A), which is also consistent with previous studies. However, the same cannot be said of CTCF-CTCF distances, which appear indistinguishable from what is expected by a random placement of sites (Wilcoxon $p = 0.1268$) (Figure S12B).

S2.14 Predicted enhancers in the ENCODE regions are enriched near cell-type specific genes

To expand our investigation across additional cell lines, we focused on differentially expressed genes between pairs of cell lines in the ENCODE regions. We counted the number of enhancers near the differentially expressed genes in the neighboring domains defined by CTCF sites. We found that enhancers are enriched near differentially expressed genes as compared to the same genes that are differentially repressed in another cell type, and this enrichment is largely confined within CTCF binding sites that directly flank the gene's TSS (Figure S13B). On average within this block, there are 0.82 enhancers per differentially down-regulated gene, while there are 1.83 enhancers per differentially up-regulated gene (Figure S13C). This 2.2-fold difference suggests that cell-type specific expression is influenced by enhancers and that the action of enhancers is distance-dependent and favoring proximal promoters. When we focused only on the enhancer closest to the differentially expressed gene rather than all enhancers within a CTCF block, we find smaller difference between the distributions of enhancers in up- and down-regulated genes (Figure S13D). The smaller 1.76-fold difference observed here further emphasizes that multiple enhancers, and not just the single closest enhancer, are likely required to regulate differential gene expression of a single promoter.

S2.15 Histone modification-based prediction of promoters on a genome-scale

Using the genome-wide ChIP-chip enrichment profiles of H3K4me1 and H3K4me3, we used our histone modification-based prediction method to make 13116 promoter predictions (Figure S10A). We found that 9835 (75%) predicted promoters overlap with 5'-ends of UCSC Known Genes[29] (Figure S10B). We also compared the promoter predictions to the RIKEN human CAGE data set[30] and observed that 11001 (83.9%) overlap with multiple CAGE tags. Further, our prediction model correctly located 76% of active RefSeq transcription start sites (TSS) (Figure S10C) and even 31.5% of inactive TSS, consistent with recent studies demonstrating the presence of similar chromatin landmarks at most promoters in the human genome[31]. We also examined the overlap of predicted promoters with CpG islands (as annotated at the UCSC genome browser[32]), sequence elements conventionally understood to be associated with many promoters[13]. The vast majority of promoter predictions (11186, 85.1%) overlap CpG islands, representing almost half (43.3%) of the genome's CpG islands (Figure S9D). These findings agree with our previous genome-wide promoter analysis[2] and are comparable to the specificity and sensitivity of our prediction model in the ENCODE regions[13].

S2.16 P-values

The p-values for correlations were obtained from using the Matlab corr function. This p-value measures the probability that there is no correlation between the two variables, against the alternative that the correlation is non-zero. The p-values for Wilcoxon rank sum tests were obtained from the Matlab ranksum function.

S2 Supplemental figure/table captions

FIGURE CAPTIONS

Figure S1: Quantitative comparison of regulatory elements in the ENCODE regions

Pearson correlation coefficients for ChIP-chip enrichment profiles (see Supplementary Information) calculated between cell types for (A) chromatin modifications at promoters, (B) CTCF binding at putative insulators, (C) p300 binding and chromatin modifications at putative enhancers, and (D) chromatin modifications at predicted enhancers. (E) We performed k-means clustering on CTCF enrichment at 729 non-redundant CTCF binding sites found by Mpeak. For comparison we have also shown the enrichment patterns from a genome-wide study in IMR90 cells[3], which supports the cell type-invariant nature of CTCF binding. (F) We clustered the chromatin modifications at 411 non-redundant p300 binding sites in HeLa, GM, and K562 cells. Enrichment of p300 binding, which was not a criteria in the clustering, confirms the cell type-specificity of the chromatin marks at enhancers.

Figure S2: ChIP-chip enrichment profiles at Gencode promoters in the ENCODE regions

ChIP-chip enrichment data for all marks in all cell types are displayed as 10 kb windows centered on the TSS of all Gencode genes in the ENCODE regions.

Figure S3: Verification of histone modification-based prediction of enhancers in the ENCODE regions

(A-D) The percentage of predicted enhancers within 2.5 kb of hypersensitive sites in HeLa, GM, K562, and ES cells as previously defined[23]. (E-G) The percentage of p300 sites mapped in HeLa, GM, and K562 cell lines within 2.5 kb of predicted enhancers (see Supplementary Information).

Figure S4: Predicted ES enhancers are enriched in known ES-specific transcription factors

Known NANOG, OCT4, and SOX2 binding sites at predicted enhancers in the ENCODE regions (see Supplementary Information).

Figure S5: Comparison of cell type-specific gene expression in four cell types

We used Shannon entropy (see Supplementary Information) to rank genes by the specificity of their expression levels in HeLa as compared to three other cell lines (K562, GM06990, and IMR90 cells, representing leukemia, lymphoblast, and fibroblast lineages, respectively). The most HeLa-specific expressed genes are found at the top of the cluster (represented in red in Figure 3B), while genes that are specifically repressed in HeLa cells are found at the bottom (green in Figure 3B). Genes in the middle portion of the cluster (black in Figure 3B) have expression levels that are similar in all four cell lines.

Figure S6: CTCF enrichment at genome-wide putative insulators in IMR90 cells

Experimentally-determined binding sites published for CTCF in IMR90, HeLa, and CD4+ T cells[3,25,26] were combined into one set of binding sites, and the ChIP-chip enrichment data at all of these sites from IMR90 cells are visualized as 10 kb windows centered at the CTCF binding sites as described above, organized by genomic position as in Figure 2A. These data support the consistency of CTCF binding across cell types.

Figure S7: Cell type-specificity of predicted enhancer activity in reporter assays

In addition to the HeLa-specific enhancers and random regions assayed in Figure 2F, additional K562-specific enhancers were cloned and assayed for reporter activity in HeLa cells. Enhancers predicted specifically in K562 cells (blue bars) were much less likely to be active in HeLa cells than the HeLa-specific enhancers (red), and the median activity is not significantly different from random regions (gray). The dashed line represents a significance threshold of $p = 0.01$ as in Figure 2F.

Figure S8: Summary of ChIP-chip and expression experiments.

The number of biological replicates for each cell-type and antibody pair.

Figure S9: Chromatin acetylation features at promoters and enhancers in the ENCODE regions.

ChIP-chip was performed on the acetylated histones H3K9Ac, H3K18Ac, and H3K27Ac, and the enrichment was compared to the (A) promoter and (B) p300 clusters from Heintzman et al in HeLa cells [13]. Each horizontal line details the ChIP-chip enrichment of various chromatin modifications and transcription factors in 10 kb windows. For consistency in comparison, we clustered the data in the same order as Heintzman et al.[13], which used k-means clustering. All three active promoter clusters P2, P3, and P4 are highly enriched in all three acetylated histones, whereas the enhancer clusters are mostly enriched in H3K18Ac and H3K27Ac, but have only weak H3K9Ac enrichment. Average profiles of log enrichment ratios for promoters or p300 binding sites in each cluster are shown at the bottom of each panel.

Figure S10: Active promoter predictions in the human genome.

(A) We predict 13116 active promoters in HeLa cells based on chromatin signatures for H3K4me1 and H3K4me3 as determined by ChIP-chip using genome-wide tiling microarrays. (B) 75% of promoter predictions map to 5' ends of UCSC Known Genes, indicating a high degree of specificity. (C) 76% of active promoters (defined as RefSeq TSS for expressed transcripts) are correctly predicted, indicating a high degree of sensitivity. (D) 85.1% of promoter predictions overlap with CpG islands (defined by UCSC Genome Browser), accounting for close to half of the CpG islands in the genome.

Figure S11: Chromatin features at CTCF binding sites.

Heat-map of the chromatin (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K18ac, H3K27ac) and transcriptional (p300, TAF1, CTCF) features within a 10-kb window of 729 consensus CTCF binding sites obtained from merging CTCF sites called for the five cell types in the ENCODE regions. This heat-map is the result of performing k-means clustering (k=4).

Figure S12: Distribution of promoters and insulators in the ENCODE regions.

(A) The distribution of adjacent TSS-TSS distances (red) for Gencode TSSs, as compared to a random placement of sites (blue). (B) The distribution of adjacent CTCF-CTCF distances (red), as compared to a random placement of sites (blue).

Figure S13: ENCODE Enhancers are clustered at differentially expressed genes.

(A) To show that enhancers are clustered, we compute the distance between adjacent enhancers and examine the distribution of these distances. The distribution of adjacent enhancer-enhancer distances (red), as compared to 1000 sets of randomly placed sites (blue), indicates that enhancers are highly clustered. (B) A CTCF block is defined by flanking CTCF binding sites. Using the 729 consensus CTCF binding sites to define CTCF blocks, we count the average number of enhancers found in blocks relative to the TSSs of differentially expressed and repressed genes. For a given TSS, CTCF block 0 is defined by the CTCF binding sites immediately flanking the TSS, CTCF block -1 is the block immediately upstream of CTCF block 0, CTCF block +1 is the block immediately downstream of CTCF block 0, etc. Differentially expressed genes are enriched in enhancers when compared to differentially repressed genes, with the strongest enrichment found in CTCF block 0.The dotted line indicates the expected average number of enhancers in a CTCF block. For HeLa, GM, and K562, differential expression is defined by an RMA p-value cutoff of 0.01 and a fold change cutoff of 2.0. (C) A detailed view of the distribution of enhancers in CTCF block 0. Here, we show the distribution of enhancer-TSS distances all enhancers within this CTCF block. Negative distances indicate upstream enhancers, while positive distances indicate downstream enhancers. Enhancers are more concentrated to differentially expressed genes relative to differentially repressed genes. (D) Rather than examining the distribution of all enhancer-TSS distances in a differentially expressed/repressed gene's CTCF block, we examined only the closest one here. While we do observe enrichment in differentially expressed genes, the effect is smaller than that observed when we consider all enhancer-TSS distances.

Figure S14: CTCF sites may serve as domain boundaries for promoter-enhancer interactions

Insulators bound by CTCF are thought to block promoter-enhancer interactions that would otherwise occur in the absence of CTCF (upper panel), a model supported by the enrichment of predicted enhancers in domains adjacent to genes that are upregulated in response to CTCF-depletion by siRNA (lower panel, red bars). Enhancers are depleted in domains adjacent to downregulated genes (green bars) relative to unchanged genes (black bars) and a random distribution (gray lines). Gene expression data are from a recently published study[25]. Domains are defined as the regions between CTCF sites as recently reported[3,25,26]; enhancers were counted in the five domains adjacent to each gene, upstream and downstream, and summed across respective domains to calculate enrichment relative to a random distribution.

TABLE CAPTIONS

Table S1: CTCF binding sites in five cell types in the ENCODE regions

Coordinates are listed in hg17 for 729 non-redundant CTCF binding sites identified in HeLa, GM, K562, ES, and dES cells (see Supplementary Information).

Table S2: p300 binding sites in HeLa cells in the ENCODE regions

Coordinates are listed in hg17 for p300 binding sites identified in HeLa cells (see Supplementary Information).

Table S3: p300 binding sites in GM cells in the ENCODE regions

Coordinates are listed in hg17 for p300 binding sites identified in GM cells (see Supplementary Information).

Table S4: p300 binding sites in K562 cells in the ENCODE regions

Coordinates are listed in hg17 for p300 binding sites identified in K562 cells (see Supplementary Information).

Table S5: Predicted enhancers in HeLa cells in the ENCODE regions

Coordinates are listed in hg17 for enhancers predicted in HeLa cells based on chromatin signatures (see Supplementary Information).

Table S6: Predicted enhancers in GM cells in the ENCODE regions

Coordinates are listed in hg17 for enhancers predicted in GM cells based on chromatin signatures (see Supplementary Information).

Table S7: Predicted enhancers in K562 cells in the ENCODE regions

Coordinates are listed in hg17 for enhancers predicted in K562 cells based on chromatin signatures (see Supplementary Information).

Table S8: Predicted enhancers in ES cells in the ENCODE regions

Coordinates are listed in hg17 for enhancers predicted in ES cells based on chromatin signatures (see Supplementary Information).

Table S9: Predicted enhancers in dES cells in the ENCODE regions

Coordinates are listed in hg17 for enhancers predicted in dES cells based on chromatin signatures (see Supplementary Information).

Table S10: Genome-wide predicted enhancers in HeLa cells

Coordinates are listed in hg17 for 36589 enhancers predicted in HeLa cells based on chromatin signatures (see Supplementary Information).

Table S11: Clone information for reporter assays

Coordinates (hg17) and primers used to amplify regions containing predicted enhancers in HeLa (H1-H9) and K562 (K1-K9) cells for cloning and reporter assays, as well as random regions selected as controls (R1-R10).

Table S12: Known motifs in predicted enhancers

Enrichment of motifs in enhancers was analyzed as described [11,12]. Over-conservation and Enrichment are calculated as the excess conservation and overabundance, respectively, of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. All significance values are expressed as Z-scores, corresponding to the number of standard deviations away from the mean of a normal distribution.

Table S13: De novo motifs enriched in predicted enhancer regions

Known Match score represents the shared information content between novel and known motif[12]. Over-conservation is calculated as the excess conservation of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. Enrichment is calculated as the over-abundance of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. Enhancer-specific motifs are those lacking significant promoter enrichment. All significance values are expressed as Z-scores, corresponding to the number of standard deviations away from the mean of a normal distribution.

Table S14: Genome-wide predicted enhancers in K562 cells

Coordinates are listed in hg17 for 24566 enhancers predicted in K562 cells based on chromatin signatures (see Supplementary Information).

Table S15: Overlap of predicted enhancers in HeLa with transcription factor binding sites in other cell types

Coordinates are listed in hg17 for each HeLa predicted enhancer with notation of overlap with experimentally determined transcription factor binding sites (see Supplementary Information).

Table S16: STAT1 binding sites in the genome of IFNγ-treated HeLa cells

Coordinates are listed in hg17 for 1969 STAT binding sites as determined by ChIP-chip.

REFERENCES

[1] Ludwig, T. E. et al., Feeder-independent culture of human embryonic stem cells. *Nat Methods* **3** (8), 637 (2006).

[2] Kim, T. H. et al., A high-resolution map of active promoters in the human genome. *Nature* **436** (7052), 876 (2005).

[3] Kim, T. H. et al., Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128** (6), 1231 (2007).

[4] Heintzman, N. D. et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39** (3), 311 (2007).

[5] Crawford, G. E. et al., DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* **3** (7), 503 (2006).

[6] Birney, E. et al., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447** (7146), 799 (2007).

[7] Song, J. S. et al., Model-based analysis of two-color arrays (MA2C). *Genome Biol* **8** (8), R178 (2007).

[8] Schug, J. et al., Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* **6** (4), R33 (2005).

[9] Felsenstein, J. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle* (2005).

[10] Xie, X. et al., Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434** (7031), 338 (2005).

[11] Kheradpour, P., Stark, A., Roy, S., and Kellis, M., Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* **17** (12), 1919 (2007).

[12] Stark, A. et al., Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450** (7167), 219 (2007).

[13] Heintzman, N. D. et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39** (3), 311 (2007).

[14] Wu, Z. and Irizarry, R. A., Preprocessing of oligonucleotide array data. *Nat Biotechnol* **22** (6), 656 (2004).

[15] Hatzis, P. and Talianidis, I., Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol Cell* **10** (6), 1467 (2002).

[16] Agalioti, T. et al., Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* **103** (4), 667 (2000).

[17] Pokholok, D. K. et al., Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122** (4), 517 (2005).

[18] Liu, C. L. et al., Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* **3** (10), e328 (2005).

[19] Zheng, M., Barrera, L. O., Ren, B., and Wu, Y. N., ChIP-chip: data, model, and analysis. *Biometrics* **63** (3), 787 (2007).

[20] Koch, C. M. et al., The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17** (6), 691 (2007).

[21] Gaszner, M. and Felsenfeld, G., Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7** (9), 703 (2006).

[22] Wei, G. H., Liu, D. P., and Liang, C. C., Chromatin domain boundaries: insulators and beyond. *Cell Res* **15** (4), 292 (2005).

[23] Xi, H. et al., Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* **3** (8), e136 (2007).

[24] Boyer, L. A. et al., Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122** (6), 947 (2005).

[25] Wendt, K. S. et al., Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451** (7180), 796 (2008).

[26] Barski, A. et al., High-resolution profiling of histone methylations in the human genome. *Cell* **129** (4), 823 (2007).

[27] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B., Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316** (5830), 1497 (2007).

[28] Berman, B. P. et al., Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* **99** (2), 757 (2002).

[29] Hsu, F. et al., The UCSC Known Genes. *Bioinformatics* **22** (9), 1036 (2006).

[30] Carninci, P. et al., Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38** (6), 626 (2006).

[31] Guenther, M. G. et al., A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130** (1), 77 (2007).

[32] Kent, W. J. et al., The human genome browser at UCSC. *Genome Res* **12** (6), 996 (2002).