Supplementary information, Data S1

Supplemental Figures, Tables, Legends and Procedures for Hawkins, Hon *et al.*

Figure S1, related to Figure 1. k-means cluster of modifications at 22,047 gene transcription start sites (center of 10-kb window). The figure is organized as in **Figure 1**. k-means = 4.

Figure S2, related to Figure 2. (A) k-means cluster of predicted enhancers in hESCs and DFCs. k-means = 4. (B) Distribution of distances between adjacent enhancers.

Figure S3, related to Figure 2. Clustering of histone modifications at distal SOX2 and NANOG binding sites not predicted as enhancers. Binding sites are ranked by acetylation levels. The clusters illustrate that most sites present some H3K4me1 and H3K27ac, suggesting they are enhancer sites not called by the prediction algorithm. SOX2 sites are shown in the left cluster, NANOG sites are in the right cluster.

Figure S4, related to Figure 3. Additional analysis of CTCF binding site analysis and CTCF-defined domains
(A) Clustergram at 29,880 combined CTCF binding sites determined from IMR90 (Kim et al, 2007), HeLa (Wendt et al., 2008), and CD4 cells (Barski et al., 2007), and assessed for in genome-wide CTCF binding sites found in IMR90 and hES cells.
(B) k-means clustergram of CTCF binding sites in the ENCODE region from HeLa, GM06990 (GM), K562 leukemic cells, hESCs, and DFCs from Heintzman et al., 2009. These are compared to ENCODE regions extracted from genome-wide data in IMR90 cells (Kim et al., 2007) and hESCs presented here (blue).
(C) The genome-wide distribution of adjacent CTCF-CTCF distances.
(D) The genome-wide distribution of the number of promoters per CORD.
(E) The genome-wide distribution of adjacent enhancer-enhancer distances for: (top) hESC predicted enhancers only, (middle) DFC predicted enhancers only, and (bottom) combined predictions.
(F) Reporter assays of enhancer function at 10 predicted DFC-specific enhancers, 2 randomly chosen genomic regions as negative controls, and 2 hESC-specific enhancers as positive controls, cloned downstream of a luciferase gene and tested for activity in hESCs. The dashed red line indicates a p-value cutoff of 1%.

Figure S5, related to Figure 3. Changes in enhancer numbers correspond to changes in gene expression
(A) For each of the 1000 hESC-specific (red), 1000 DFC-specific (green), and 1000 non-specific (black) genes, we counted the number of enhancers found before and after differentiation within the CTCF-defined domain. We then plotted the distribution of this difference normalized over the distribution of all genes.
(B) As in **A**, except directly comparing enhancer numbers within CTCF-defined domains for hESC-specific genes *(right)* and DFC-specific genes *(left)*. The enrichment ratio, as above, is shown as a heat-map (red = enrichment; green lack of enrichment). **Figures 3A, 3B** show a transition in enhancer use

for cell-type specific gene expression
(C) Plot of differential enhancer number as a function of differential expression for all 22,047 genes, averaged into 100 gene bins. The fitted line shows a correlation between differential enhancer number and differential gene expression.

Figure S6, related to Figure 3. UCSC Genome browser shots of histone modifications and enhancer predictions in CTCF-defined blocks (CORDs) containing (A) *SOX2*, (B) *OCT4(POU5F1)*, (C) *NANOG*, and the (D) *HOXB* locus. Figure is displayed as those in the main text. Gene names are at the 5' end of the gene.

Figure S7, related to Figure 4. Shared Enhancers May be Poised
(A) Clustering of 8863 predicted enhancers that are shared between hES and DF cells. The clustering was ranked on H3K27ac intensity. Each end of the spectrum shows that some enhancers exhibit cell type-specific acetylation, although mono-methylated in both cell types.
(B) Assessment of enhancer enrichment during a time course of gene expression during BMP4 treatment. Early response genes are more enriched in shared enhancers acetylated in DFCs (yellow) compared to DFC-specific (red), hESC-specific (dark blue), or shared enhancers with hESC acetylation (light blue).

Table S1, related to Figure 1. Figure 1 Sorted Gene List and Expression Values Ranked by Cg

Table S2, related to Figure 1. Top 1% genes exhibiting H3K27 switch following differentiation.

Table S3, related to Figure 2. Enhancers Predicted by Chromatin Signatures in hESC and DFC..

Table S4, related to Figure 2. Complete Motif List for Enhancers.

Table S5, related to Figure 3. CTCF Bound Sites

Table S6, related to Figure 3. Genes within CTCF-defined domains with top enhancers (top 1% Ce ranking)

Table S7, related to Figure 4. Genes Up-regulated at 3hrs with Poised Enhancer

SUPPLEMENTAL EXPERIMENTAL PROCEDURES:

Additional Methods Related to Main Figures

Figure 1

We have expression data for 22047 genes. For each gene and cell type, we calculate the sum of the log2 enrichment of H3K27ac and H3K27me3 in a 10-kb window centered at the TSS, and take the difference H3K27ac – H3K27me3 representing the enrichment of H3K27ac over H3K27me3 in a single cell type. We then compute the difference of this value over the 2 different cell types (DFC – hESC), and rank all genes using this difference. Negative differences indicate ES-specific H3K27ac and DFC-specific H3K27me3. Positive differences indicate ES-specific H3K27me3 and DFC-specific H3K27ac. See gene expression below.

Figure 2

A) We combined all hESC and DFC enhancers, merging those that were within 2.5 kb of each other, to get a list of 53374 unique enhancers. We then employed the same method as in Figure 1A, except we only use H3K27ac in the ranking.

C) The distribution of enhancers at 5', 3', intragenic, and intergenic regions. Using the UCSC known genes (070507) definition, we first found enhancers within 2.5 kb to 5' ends. Of those enhancers not recovered, we then found those within 2.5 kb to 3' ends. Of those remaining, we found those within known gene 5' and 3' ends. Those enhancers not recovered are intergenic. For comparison, we randomly placed 1000000 sites onto places well-represented on the Affymetrix arrays and computed the distribution of these sites in the same manner.

Figure 3

B) Enhancers are localized to cell type-specific genes, and this localization is bounded by the promoter's flanking CTCF sites. We partitioned the enhancer heatmap in Figure 2a into thirds: hES-specific H3K27ac enhancers in the top third, non-specific H3K27ac in the middle third, and DFC-specific H3K27ac in the bottom third. To examine cell-type specificity, we focused on 3 sets of genes: the 1000 most ES-specifically expressed genes, the middle 1000 non-specifically expressed genes, and the 1000 most DFC-specifically expressed genes. For each category of enhancer and gene, we computed the average number of enhancers in various CTCF blocks around the genes' promoters.

C) As in Figure 3E, but focusing on the distance distribution of enhancers at CTCF block 0.

D) H1 cells were grown on Matrigel with MEF conditioned medium plus bFGF (100ng/ml). Enhancers were cloned as previously described (Heintzman et al., 2009). Primers are available upon request. Briefly, ~1-3kb genomic regions were PCR amplified with primers containing linkers suitable for ligation-independent cloning utilizing the Infusion Dry-Down PCR Cloning system (Takara Bio, Clontech). Enhancer amplicons were cloned downstream of the luciferase gene in the pGL3 Promoter vector (Promega). Loci were selected based on importance to ES cell biology for gene within the same CTCF-block. Transfections and reporter assays: Transfections were carried out in triplicate in 96-well format. H1 cells were plated and allowed to reach 70-80% confluency prior to transfection, usually 2 days. Cells were transfected using FuGENE HD transfection reagent (Roche Diagnostics) at a ratio of 5:2 (ul FuGENE : ug DNA). Cells were transfected with 200ng of each reporter construct and 200pg of pRL-CMV (Promega) for normalization. Reporter constructs and FuGENE reagent were diluted in

DMEM and added cell medium. Following transfection, cells were incubated under normal conditions for 48 hours then lysed and assayed for reporter activity using the Dual Luciferase Reporter Assay System (Promega) and GeniosPro Luminometer (TECAN) according to manufacturer specifications.

(E) 3C was performed as described in (Hagege et al., 2007). Briefly, cross-linked nuclei were lysed with SDS/Triton-X and digested with EcoRI restriction enzyme overnight. The enzyme was inactivated with SDS/Triton-X and ligation was performed with T4 DNA ligase. Samples were then reverse cross-linked overnight at 65 C, followed by RNAse A and Proteinase K treatment, phenol chloroform extraction, and DNA precipitation. PCR was performed using the Roche LightCycler 480 in 384 well format. Primers are available upon request. Each PCR was performed in triplicate, and the outlier replicate was discarded. PCR efficiency was normalized to a PCR reaction spanning a region not cut by EcoRI.

Figure 4

A) We focus on all genes in Figure 1A along with 3 sets of enhancers: those unique to hES cells, those unique to DFCs, and those found in both cells. For each such gene, we count the number of enhancers predicted in each cell type flanked by insulator-binding CTCF sites called in ES cells. We sort the genes by differential (DFC - ES) gene expression, and use a sliding window of 1000 genes to give a profile of the average number of enhancers for each cell type as a function of average differential gene expression. This gives three profiles, one for each enhancer set. 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.

References

Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forne, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc 2, 1722-1733.

Additional Methods for Supplemental Figures

We began by mapping H3K4me1 and H3K4me3 in both hESCs and DFCs to distinguish promoters and enhancers in each genome (Heintzman et al., 2007). This was done using Affymetrix whole genome tiling arrays consisting of 7 chips at a 35 bp resolution. To make the best use of publicly available data, we used a previously published genome-wide map for H3K27me3 in hESCs (Pan et al., 2007), which used NimbleGen System's platform. We also used the NimbleGen platform to construct genome-wide maps for H3K27ac and CTCF in hESCs, and for H3K27ac and H3K27me3 in DFCs. Focusing on the ENCODE regions (1% representation of the human genome – (Consortium, 2004)), we observe that the different platforms are highly reproducible (data not shown).

Figure S5. (A) For each of the genes in the 3 categories of cell type-specific expressed genes in Figure 3C-E, we computed the difference in the number of enhancers in DFCs versus hES cells flanked by CTCF sites. We then collected these numbers to generate a distribution of enhancer number changes for each class of gene. To normalize, we compute the difference in the number of enhancers for all genes. We then define enrichment to be the ratio between the distribution of a class of genes to that of all genes.

(B) As in Figure S6A, but instead of computing the difference in enhancers, we count the number of enhancers in each cell type. Enrichment is determined as in Figure 4a.

(C) For each gene in Figure 1A, we compute the number of enhancers flanked by CTCF sites and the gene's differential expression. We sort this list by expression and average groups of 100 genes (each point), and fit a line to this data.

Figure S7. (A) Shared enhancers are ranked by Ce and is main text. (B) As in **Figure 4C**, **4D**.

Additional Methods for Main Text

Gene Expression Data Analysis for hESCs and DFCs

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.

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.

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 (1). 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

References

(1) Irizarry RA et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249 (2003).

ChIP-Seq

ChIP for SOX2 (R&D Systems, #AF2018) and NANOG (R&D Systems, #AF1997) were carried out as described above with 500ug chromatin and 5ug antibody. ChIP libraries for sequencing were prepared following Illumina protocols with the minor modifications. Following linker ligation libraries were run on 8% acrylamide gel and size selected for 175-250bp. This was repeated following PCR amplification. After each size selection, acrylamide was shreaded, and incubated with 300ul EB buffer from Qiagen PCR Purification kit o/n at 4ºC or 50 ºC for 20 mins with shaking. DNA was eluted using Nanosep MF filter tubes. Libraries were sequenced using Illumina's GAII machine following their protocols. Following sequencing cluster imaging, base calling and mapping were conducted using the Illumina pipeline. Total mapped tags were paired down to only the monoclonal unique tags. These are tags that mapped to one location in the genome and each sequence is represented once. A total of 3,829,171 monoclonal unique tags were analyzed for NANOG, and 7,663,457 for SOX2. An equal number of tags from input DNA was utilized in MACS peak finding.

Motif Discovery

Data: 637 genes down-regulated during the first 48 hours of differentiation induced by BMP4 treatment were defined as human embryonic stem cell (hESC) specific genes while 1214 genes up-regulated 48 or more hours after BMP4 treatment were defined as differentiation specific genes. 1028 enhancers identified in hESCs were mapped to the hESC specific genes bounded by insulators, and 3221 enhancers identified in BMP4 with FGF (have to specify slightly different conditions here) differentiated cells were mapped to the differentiation specific genes bounded by insulators. Genomic sequences of these hESC and differentiation specific enhancers of 5000 kbs were extracted from the UCSC GoldenPath database of the hg 17 assembly (http://hgdownload.cse.ucsc.edu/downloads.html). Two data sets with 1028 and 3221 random genomic sequences of 5000 kbs were also extracted from the same database as controls.

Procedure: 566 TRANSFAC and 96 vertebrate transcription factors (TFs) motif matrices were downloaded from the JASPAR database (http://jaspar.genereg.net/). MotifLocator, software based on a classical position-weight matrix scoring scheme, was downloaded from the INCLUSive database (http://homes.esat.kuleuven.be/~thijs/download.html) (3) and was used to search the hESC and differentiation specific enhancers for potential binding sites of the 96 TFs. The motifs' ability to classify foreground sequences from background sequences was measured by the balanced misclassification error rate (1). The error was defined as:

$$ErrorRate = 1 - [(Sensitivity + Specificity)/2]$$

Sensitivity was defined as the proportion of sequences in the foreground set containing a motif, and specificity was defined as the proportion of sequences in the background set without the motif (1). The threshold for motif matching was optimized for each matrix to minimize the error rate. To identify hESC specific TFs, hESC specific enhancers were used as foreground sequences while differentiation specific enhancers were used as background sequences. Correspondingly, to identify differentiation specific TFs, the foreground and the background data sets were flipped.

The significance of the balanced misclassification error rate for a motif (p-value) for a given comparison was determined by the distribution of the error rate. This distribution was estimated by a permutation method (1). To further verify a motif's ability to classify foreground sequences from background sequences, a 95% confidence interval (95% CI) (2) of the difference between the proportion of the sequences with the motif in the foreground set and the proportion of the sequences with the motif in the background set was calculated for each of the 96 TFs. If zero is not in the 95% CI, the difference between the two sets is significant at the 5% level. Otherwise, it is not significant. We filtered the results to include motifs with p-value <0.05, specificity >2/3 and zero being outside the 95% CI. To prove the abilities of this algorithm to identify the difference between hESC specific enhancers and differentiation specific enhancers, two random genomic data sets with 1028 and 3221 sequences were compared with each other. The difference between these two data sets was much less significant than the one between hESC and differentiation enhancers, indicating the great power of the algorithm to distinguish two data sets

Reference

(1) Barrera LO, Li Z, Smith AD, Arden KC, Cavenee WK, Zhang MQ, Green RD, Ren B. Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. 2007. *Genome Res*. 18(1):46-59.

(2) Douglas G. Altman, David Machin, Thevor N. Bryant and Martin J. Gardner. 2000. Statistics With Confidence: Confidence Intervals and Statistical Guidelines. Published by BMJ.

(3) Thijs G., Moreau Y., De Smet F., Mathys J., Lescot M., Rombauts S., Rouzé P., De Moor B., Marchal K., 2002. INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling, Bioinformatics, 18(2), 331-332.

REFERENCES FOR TABLE 2

Barkett, M., and Gilmore, T.D. (1999). Control of apoptosis by Rel/NF-kappaB transcription factors. Oncogene *18*, 6910-6924.

Cole, T.J., Blendy, J.A., Monaghan, A.P., Krieglstein, K., Schmid, W., Aguzzi, A., Fantuzzi, G., Hummler, E., Unsicker, K., and Schutz, G. (1995). Targeted disruption of the glucocorticoid receptor gene blocks adrenergic chromaffin cell development and severely retards lung maturation. Genes Dev *9*, 1608-1621.

Drolet, D.W., Scully, K.M., Simmons, D.M., Wegner, M., Chu, K.T., Swanson, L.W., and Rosenfeld, M.G. (1991). TEF, a transcription factor expressed specifically in the anterior pituitary during embryogenesis, defines a new class of leucine zipper proteins. Genes Dev *5*, 1739-1753.

Hennighausen, L., and Robinson, G.W. (2008). Interpretation of cytokine signaling through the transcription factors STAT5A and STAT5B. Genes Dev *22*, 711-721.

Hock, H., Hamblen, M.J., Rooke, H.M., Schindler, J.W., Saleque, S., Fujiwara, Y., and Orkin, S.H. (2004). Gfi-1 restricts proliferation and preserves functional integrity of haematopoietic stem cells. Nature *431*, 1002-1007.

Holtschke, T., Lohler, J., Kanno, Y., Fehr, T., Giese, N., Rosenbauer, F., Lou, J., Knobeloch, K.P., Gabriele, L., Waring, J.F., *et al.* (1996). Immunodeficiency and chronic myelogenous leukemia-like syndrome in mice with a targeted mutation of the ICSBP gene. Cell *87*, 307-317.

Kinoshita, K., Ura, H., Akagi, T., Usuda, M., Koide, H., and Yokota, T. (2007). GABPalpha regulates Oct-3/4 expression in mouse embryonic stem cells. Biochem Biophys Res Commun *353*, 686-691.

Lehmann, O.J., Sowden, J.C., Carlsson, P., Jordan, T., and Bhattacharya, S.S. (2003). Fox's in development and disease. Trends Genet *19*, 339-344.

Min, I.M., Pietramaggiori, G., Kim, F.S., Passegue, E., Stevenson, K.E., and Wagers, A.J. (2008). The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. Cell Stem Cell *2*, 380-391.

Mohun, T., and Sparrow, D. (1997). Early steps in vertebrate cardiogenesis. Curr Opin Genet Dev *7*, 628-633.

Nerlov, C. (2007). The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. Trends Cell Biol *17*, 318-324.

Nishimura, G., Manabe, I., Tsushima, K., Fujiu, K., Oishi, Y., Imai, Y., Maemura, K., Miyagishi, M., Higashi, Y., Kondoh, H., *et al.* (2006). DeltaEF1 mediates TGF-beta signaling in vascular smooth muscle cell differentiation. Dev Cell *11*, 93-104.

Perrotti, D., Melotti, P., Skorski, T., Casella, I., Peschle, C., and Calabretta, B. (1995). Overexpression of the zinc finger protein MZF1 inhibits hematopoietic development from embryonic stem cells: correlation with negative regulation of CD34 and c-myb promoter activity. Mol Cell Biol *15*, 6075-6087.

Rosfjord, E., and Rizzino, A. (1994). The octamer motif present in the Rex-1 promoter binds Oct-1 and Oct-3 expressed by EC cells and ES cells. Biochem Biophys Res Commun *203*, 1795-1802.

Rudnicki, M.A., and Jaenisch, R. (1995). The MyoD family of transcription factors and skeletal myogenesis. Bioessays *17*, 203-209.

Shiratori, H., Sakuma, R., Watanabe, M., Hashiguchi, H., Mochida, K., Sakai, Y., Nishino, J., Saijoh, Y., Whitman, M., and Hamada, H. (2001). Two-step regulation of left-right asymmetric expression of Pitx2: initiation by nodal signaling and maintenance by Nkx2. Mol Cell *7*, 137-149.

Smith, J. (1999). T-box genes: what they do and how they do it. Trends Genet *15*, 154-158.

Wang, Z., Wang, D.Z., Hockemeyer, D., McAnally, J., Nordheim, A., and Olson, E.N. (2004). Myocardin and ternary complex factors compete for SRF to control smooth muscle gene expression. Nature *428*, 185-189.

Withington, S.L., Scott, A.N., Saunders, D.N., Lopes Floro, K., Preis, J.I., Michalicek, J., Maclean, K., Sparrow, D.B., Barbera, J.P., and Dunwoodie, S.L. (2006). Loss of Cited2 affects trophoblast formation and vascularization of the mouse placenta. Dev Biol *294*, 67-82.

Wu da, Y., and Yao, Z. (2005). Isolation and characterization of the murine Nanog gene promoter. Cell Res *15*, 317-324.