Supporting Information

Creyghton et al. 10.1073/pnas.1016071107

SI Materials and Methods

Cell Growth and Culture Conditions. Blastocysts were isolated from the oviducts of hormone-primed [pregnant mare serum at 5 IU/mouse (catalog no. 367222; Calbiochem), human chorionic gonadotrophin at 50 IU/mL (catalog no. 230734; Calbiochem)] B6D2F1 mice (Taconic) either 3 d after fertilization or 1 d after fertilization followed by 3 d of culture in potassium simplex optimized media (Specialty Media).

C57/BL6-129JAE (v6.5) murine ES cells and J21 (C57/BL6) ES cells were grown in DMEM supplemented with 15% FBS (HyClone), 1,000 U/mL leukemia inhibitory factor 0.001% β-mercapto-ethanol (M7522; Sigma), 100 μM nonessential amino acids (11140-050; Invitrogen), 2 mM L-glutamine (25030-081; Invitrogen), 100 U/mL penicillin, and 100 μg/mL streptomycin (15140-122; Invitrogen).

Adult tissues were isolated from 4- to 6-wk-old 4f2a mice (1). CD19$^+$ proB cells were isolated by magnetic affinity cell sorting cell separation (Miltenyi Biotech) according to the manufacturer's protocol. Purified B cell subsets were resuspended in Iscove's Modified Dulbecco's Medium with 15% FCS as well as IL-4, IL-7, and stem cell factor (10 ng ml^{-1} each; Peprotech) and plated on OP9 bone marrow stromal cells (ATCC). Cells were fixed in formaldehyde or TRIzol reagent for ChIP and RNA isolation. For isolation of liver cells, mice were first perfused with 50 mL HBSS buffer (without Ca^{2+} and Mg^{2+}), then with 50 mL HBSS (without Ca^{2+} and Mg^{2+}) containing collagenase (type IV) (Sigma) (100 U mL^{-1}). Liver was dissected away from surrounding tissues and dissociated in 10 mL DAG media [phenol-red free DMEM (Gibco) and BSA 1 g per 0.5 L] and filtered two times through a sterile 100-μm cell strainer. Liver cell preparations were centrifuged at 30 × *g* for 3 min at 4 °C, and the cells were washed two times in PBS. Neural progenitors were derived by plating 4-d-old embryoid bodies and selection for 5–7 d in chemically defined ITSFn media (insulin, transferrin, selenium, and fibronectin) (2). Neural precursor cells were isolated and cultured on polyornithine-coated cell culture dishes in N3 media containing FGF and EGF (R&D Systems) as described previously (3). In the presence of growth factors these cells can be labeled homogenously with antibodies against nestin, Sox2, and A2B5. Upon growth factor withdrawal, the cells differentiate into TUJ1-positive neurons, 04-positive oligodendrocytes, and GFAP-positive astrocytes, the three major cell types of the nervous system (3), indicating that the precursor cells sustain differentiation potential.

RNA Isolation and Microarray Analysis. RNA was isolated using TRIzol reagent (Invitrogen) according to the manufacturer's protocol and DNase treated using the DNA-Free RNA kit (R1028; Zymo Research). Cy3 dye-labeled cRNA samples were prepared using Agilent's QuickAmp sample labeling kit. Input was 0.5 μg total RNA. Briefly, first- and second-strand cDNA are generated using Moloney Murine Leukemia Virus–reverse transcriptase enzyme and an oligo-dT based primer. In vitro transcription is performed using T7 RNA polymerase and either cyanine 3-CTP or cyanine 5-CTP, creating a direct incorporation of dye into the cRNA. Agilent (mouse 4x44k) expression arrays were hybridized according to our laboratories method, which differs slightly from the Agilent standard hybridization protocol. The hybridization mixture consisted of 1.65 μg cy3 dye-labeled cRNA for each sample, Agilent hybridization blocking components, and fragmentation buffer. The hybridization mixtures were fragmented at 60 °C for 30 min, followed by addition of Agilent 2X hybridization. The arrays were then hybridized for 16 h at 60 °C in an Agilent rotor incubator set at maximum speed. Arrays were washed (6× SSPE/0.005% *N*-laurylsarcosine) on a rotating platform at room temperature for 2 min, and then washed again (0.06× SSPE) for 2 min at room temperature. The arrays were then dipped briefly in acetonitrile, followed by 30 s in Agilent Stabilization and Drying Solution. Arrays were scanned with an Agilent scanner using the Agilent feature extraction software.

Chromatin Immune Precipitation. Cells were chemically cross-linked either on plate or in suspension by the addition of one-tenth volume of fresh 11% formaldehyde solution containing [1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, and 50 mM Hepes-KOH (pH 7.5)] for 15 min at room temperature. Cells were rinsed twice with 1× PBS, harvested using a silicon scraper or centrifuge, flash-frozen in liquid nitrogen, and stored at −80 °C before use. Typically 1–3 × 10^8 cells were resuspended in 10 mL lysis buffer [50 mM Hepes-KOH (pH 7.5), 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% Nonidet P-40, and 0.25% Triton X-100] and rocked at 4 °C for 10 min. Cells were pelleted at 2880 × *g* on a tabletop centrifuge at 4 °C and resuspended in wash buffer [200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, and 10 mM Tris (pH 8.0)] and rocked at room temperature for 10 min. Cells were pelleted by spinning at 2880 × *g*/4 °C for 5 min and resuspended in 3 mL of sonication buffer [1 mM EDTA, 0.5 mM EGTA, 10 mM Tris (pH 8.0), 100 mM NaCl, 0.1% Na-deoxycholate, and 0.5% *N*-lauroylsarcosine]. We used a Misonix Sonicator 3000 and sonicated at ≈20 W for 8 × 30-s pulses (60-s pause between pulses) at 4 °C for ES cells while samples were immersed in an ice bath. For adult tissues we sonicated six to seven rounds of 30-s pulses. Triton X-100 was added to the resulting whole-cell extract (1% end concentration), which was then cleared by centrifugation (20,800 × *g* at 4 °C), and supernatant was incubated overnight at 4 °C with 100 μL of Dynal Protein G magnetic beads that had been preincubated with 10 μg of the appropriate antibody for at least 3 h. Beads were washed five times with RIPA buffer [50 mM Hepes (pH 7.6), 1 mM EDTA, 0.7% deoxycholate, 1% Nonidet P-40, and 0.5 M LiCl] and once with Tris/EDTA buffer (TE) containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65 °C with occasional vortexing in elution buffer [50 mM Tris (pH 8), 10 mM EDTA, and 1% SDS], and cross-linking was reversed by overnight incubation at 65 °C. After removal of the beads, immunoprecipitated DNA was diluted 1:1 with TE and then treated with RNase A (0.2 μg/μL final) for 2 h at 37 °C, followed by proteinase K (0.2 μg/μL final) treatment for 2 h at 50 °C. DNA was purified using two consecutive phenol:chloroform extractions using Phase Lock Gel tubes (5′) and once using Qiagen PCR purification columns. The resulting DNA was either used for gene-specific PCR or further treated for analysis on the Solexa sequencer (GA2X genome sequencer) using the ChIP seq sample prep kit (1003473; Illumina) according to the manufacturer's protocol (11257047; Illumina), selecting library fragments between 200 and 350 bp. Samples were run by the Massachusetts Institute of Technology biopolymers facility (http://web.mit.edu/ki/facilities/biopolymers/index.html) using the GA2X genome sequencer (SCS v2.6, pipeline 1.5).

Data Analysis. Images that were acquired from the Illumina/Solexa sequencer were processed using the bundled Solexa image extraction pipeline [version 1.5 or 1.6 (Cassava)]. Sequences (36 bp) were aligned using MAQ software (http://maq.sourceforge.net) using murine genome National Center for Biotechnology Information Build 36 and 37 [University of California, Santa Cruz (UCSC) mm8] as the reference genome. Sequences were mapped using iterative mapping from 36 to 26 bp, excluding reads with more than one mismatch and reads that have a sequence mapping quality Phred score of <50. Reads that had more than two exact matches were also excluded to correct for sequence bias. As a minimal requirement between 5 and 10 million reads per IP had to be successfully mapped (Table S1). Analysis of our sequence data were done on the basis of previous models (4, 5). Sequences were extended −400/+600 bp for histone marks representing histone-bound DNA length plus fragment length, or +200 bp for transcription factors, and allocated in 25-bp bins (1.05×10^8 bins total). Statistically significant enriched bins were identified using a Poissonian background model, generally with a *P* value threshold of 10^{-8} to minimize false-positive results. For enhancer identification, a *P* value of 10^{-6} was used for K4me3 to minimize false-negative promoters in our enhancer pool. We used an empirical background model (H3 for histone marks, whole-cell extracts for other factors) that require genomic bins to be enriched at least threefold above background and extended regions to be enriched at least fivefold above background to correct for nonrandom enrichment observed previously (4). Enhancer locations were defined by the peak of the enriched regions, and these regions were required to be located at least 1,000 bp away (edge to edge) from a k4me3-enriched region or a known transcriptional start site (TSS) (downloaded from the UCSC table browser; http://genome.ucsc.edu/cgi-bin/hgTables?command=start Refseq), including TSS for miRNAs (4) using Galaxy (http://main.g2.bx.psu.edu). Enhancer peaks that were within 500 bp of each other were merged into one enhancer.

Identification of DNA binding motifs on enhancers. Repeat masked conserved sequences from 200-bp regions under the top enhancer peaks (75 counts or more) were analyzed using Bioprospector (6) to identify motifs having a width from 6 to 20 nt. Motifs were matched against known transcription factor motifs in TRANSFAC, UNIPROBE (mouse), and JASPAR (core) databases using TOMTOM (7).

Gene Ontology. Gene Ontology (GO) analysis was done using GOstat (http://gostat.wehi.edu.au/cgi-bin/goStat.pl) using the mgi (mouse) GO annotation database, which contains 234,858 annotations and 35,165 gene products annotated merging GO annotations if indicating genes are inclusions (8). For enhancer locations from each cell type, we selected the closest gene either upstream or downstream and subselected genes that were exclusively associated with a close enhancer in that particular cell type (Tables S1).

ChIP-Seq cluster and meta-gene analysis. Enhancers were aligned around the peak location, and genes were aligned according to the position of the transcription start site. Around these sites a region of −4 to +4 kbp was selected and subdivided in 200-bp bins. Read density profiles were normalized to the read density per million total reads per bin around the center of the region. These regions were either left unsorted or clustered using *k*-means clustering (Euclidean distance) using Cluster 3.0 software (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm) (9). Heatmaps were generated using Java treeview (http://jtreeview.sourceforge.net/) (10). Meta-gene plots were generated by collapsing all bins and calculating the average read density per bin corrected for the total number of reads.

Correlation of ChIP-seq data. Correlations were performed on 100-bp bins spanning −4k to +4k bp of each enhancer or TSS site. The maximum value of the region in each sample is defined as the peak. The peaks for all regions from all samples were quantile normalized. The bin values of each region were then rescaled according to the normalized peak value. Then bin values across all regions from each sample are then concatenated to form a binding vector. Binding vectors from all samples were clustered by hierarchical clustering using Pearson correlation and average linkage (see *Hierarchical clustering* for more details).

eRNA heatmap. H3K4me1-, H3K27ac-, p300-, and Oct4-enriched regions were filtered to be at least 3 kb from known TSS and at least 1 kb from the edge of an H3K4me3-enriched region. The 3-kb cutoff was chosen here to ensure that eRNAs would be detected instead of transcriptional-start site RNAs (TSSA-RNAs). If the centers of an H3K4me1- and a H3K27ac-enriched region were less than 3 kb apart, it was consolidated into a single H3K27ac+H3K4me1+ enhancer rather than counted as two separate regions. If there was a p300 or Oct4 binding site within 3 kb of the H3K27ac peak or K4me1 peak, RNAs were aligned to the transcription factor binding site. If there were no nearby transcription factor binding sites, the RNA was aligned to the H3K27ac peak if present or the H3K4me1 peak otherwise. In each panel an 8-kb window is shown. The window in the short RNA panel was split into 500-bp bins, which were colored black if an RNA read was present and white if no reads were detected. Previously published RNA libraries for V6.5 ES cells were used (9). The H3K27ac and H3K4me1 regions were colored in 25-nt bins with shading corresponding to the normalized reads from the corresponding ChIP-Seq data sets. The image was produced using Matlab's imagesc function.

Hierarchical clustering. Clustering uses BioPython (http://biopython.org/) and Cluster3 module (10). Pairwise sample correlation matrix was put in place of the value heatmap. Clustering trees and heatmaps are visualized in JavaTreeView (11).

Preprocessing of gene expression data. In addition to our our microarray data, a set of 531 microarray data generated using the same platform (GPL4134) was downloaded from the Gene Expression Omnibus database, which will serve as basis for the row-wise z-score transformation described below. Microarray data were processed and normalized within the array by LOESS and across arrays by quantile normalization using the limma package (12–14). To get the standardized expression of each probe values across samples, row-wise z-scores were calculated using all array data per probe across all samples.

Analysis of gene expression of enhancer-associated genes. Probes targeting the same gene were collapsed into a gene z-score taking the median of the probe z-scores. Gene z-score values were transformed into quantile scores = (number of genes in array − rank in array)/number of genes in array × 100, such that genes with highest score were given 100, and genes with lowest score were given 0. The closest genes to enhancers were found using Galaxy (http://main.g2.bx.psu.edu/). The quantile score distributions were plotted in box-plots, and the significance of differential distribution was estimated using the Mann–Whitney *U* test on quantile scores among sample pairs. *P* values were then adjusted for multiple comparison by Bonferroni correction. For Fig. 1*C* (main text) and Fig. S3, only genes associated exclusively with enhancers in the indicated cell type were plotted. For Figs. 3*E* and 4*A* (main text) and Fig. S5, genes were mutually-exclusively classified into k4+k27+, k4−k27+, and k4+k27− groups according to the following hierarchy: genes with k4+k27+ enhancers were placed into the k4+k27+ group. The remaining genes that were associated with k4−k27+ enhancers were placed into the k4−k27+ group. After that the remaining genes with k4+ enhancers were placed into the k4+k27− group.

Enrichment of ChIP-seq-identified transcription binding sites at enhancers. Cell-specific enhancers (LT) were overlapped with transcription factor (TF) binding peaks or clusters to create the cell-specific set of TF+ enhancers (LH). The collection of all enhancers was used as the population (PT), which was overlapped with the TF peaks or clusters to get all enhancers that are TF+ (PH). The Fisher exact test was performed on a 2 × 2 contingency table using the numbers [LH, PH - LH, LT - LH, (PT - LT) - (PH - LH)]. The left-tail *P* value corresponds to the *P* value of depletion, whereas the right-tail *P* value corresponds to the *P* value of enrichment. Fold enrichment was calculated as LH/LT/(PH/PT).

1. Carey BW, Markoulaki S, Beard C, Hanna J, Jaenisch R (2010) Single-gene transgenic mouse strains for reprogramming adult somatic cells. *Nat Methods* 7:56–59.
2. Okabe S, Forsberg-Nilsson K, Spiro AC, Segal M, McKay RD (1996) Development of neuronal precursor cells and functional postmitotic neurons from embryonic stem cells in vitro. *Mech Dev* 59:89–102.
3. Conti L, et al. (2005) Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol* 3:e283.
4. Marson A, et al. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134:521–533.
5. Guenther MG, et al. (2008) Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. *Genes Dev* 22:3403–3408.
6. Liu X, Brutlag DL, Liu JS (2001) BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 127–138.
7. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. *Genome Biol* 8:R24.
8. Beissbarth T, Speed TP (2004) GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20:1464–1465.
9. Sheila AC, et al. (2008) Divergent transcription from active promoters. *Science* 322: 1804–1805.
10. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20:1453–1454.
11. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20:3246–3248.
12. Ritchie ME, et al. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 23:2700–2707.
13. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. *Methods* 31: 265–273.
14. Smyth GK (2005) Limma: Linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (Springer, New York), pp 397–420.

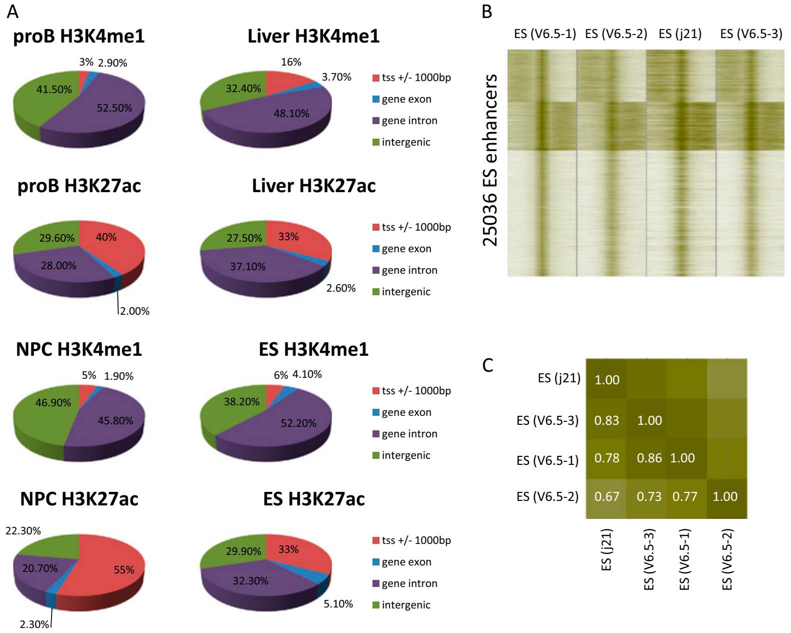

Fig. S1. Distribution of H3K4me- and H3K27ac-enriched regions across the genome. Cell-specific enhancer patterning is reproducible and not heavily dependent on genetic background. (*A*) Pie charts showing the distribution of H3K4me1- and H3K27ac-enriched regions across the genome for the indicated tissues. Distribution among TSS, gene exons, introns, and intergenic regions are displayed by the color coding. (*B*) Heatmap of distal ES-specific enhancers indicated cell/tissue types showing similar enhancer distribution. V6.5 (129JAE/C57B6) ES cell data sets (V6.5-1) are displayed from our laboratory and the Broad Institute (1) indicated as ES V6.5-2. ES V6.5-3 was published previously (2). ES J21 has a C57/B6 inbred background. Four-kilobase pairs around the enhancer peaks are displayed. (*C*) Correlation analysis for the H3K4me1 marks shown in *A*; color intensities are a measure of correlation (Pearson), which are also indicated by numbers.

[1] Meissner A, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766–770.
[2] Carey BW, et al. (2009) Reprogramming of murine and human somatic cells using a single polycistronic vector. *Proc Natl Acad Sci USA* 106:157–162.

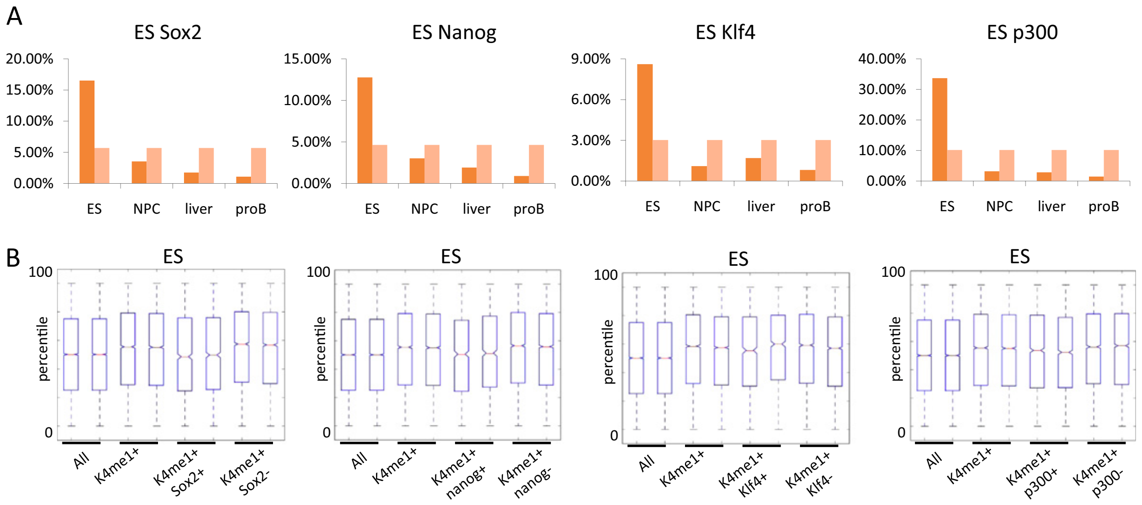

Fig. S2. Enrichment of transcription factors p300, Sox2, Klf4, and Nanog at enhancers in ES cells and their effect on proximal gene expression. (*A*) Enrichment graphs for the indicated factors at enhancers in the four cell types indicated (*x* axis). *y* axis displays percentage of cell-specific enhancers. Light bars show percentage enriched expected by chance, dark bars show actual percentage enriched. (*B*) Gene expression microarray data displayed in box-plots for genes found specifically associated with enhancers that are also bound by (+) or not bound by (–) the indicated factors. Solid bars of boxes display the 25–75% of ranked genes, with the mean indicated as an intersection. *P* values are Bonferroni corrected.

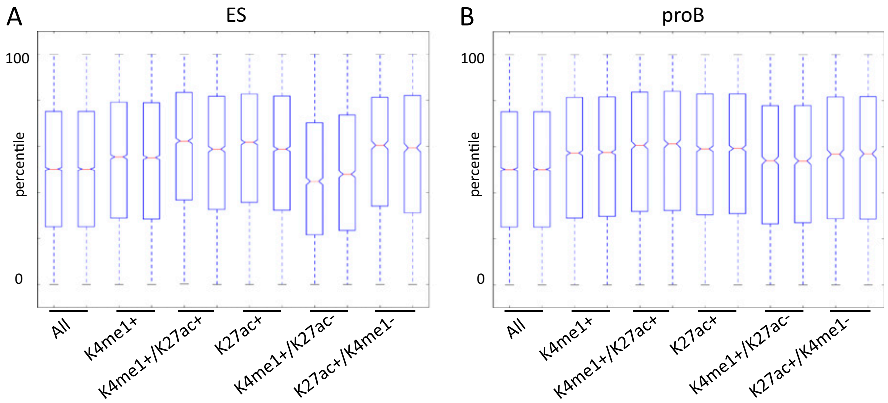

Fig. S3. Enhanced proximal gene activity of H3H3K4me1-enriched distal elements is a function of H3K27ac enrichment in different cell types. (*A* and *B*) Gene expression microarray data generated in duplicate from (*A*) ES cells or (*B*) proB cells. Box-plots show all genes (All) and genes found specifically associated with enhancers in (*A*) ES cells or (*B*) proB cells according to presence (+) or absence (–) of distal H3K4me1 or H3K27ac. A simple hierarchical model corrects for multiple enhancers being associated with single genes (*SI Materials and Methods*). Solid bars of boxes display the 25–75% of ranked genes, with the mean indicated as an intersection.

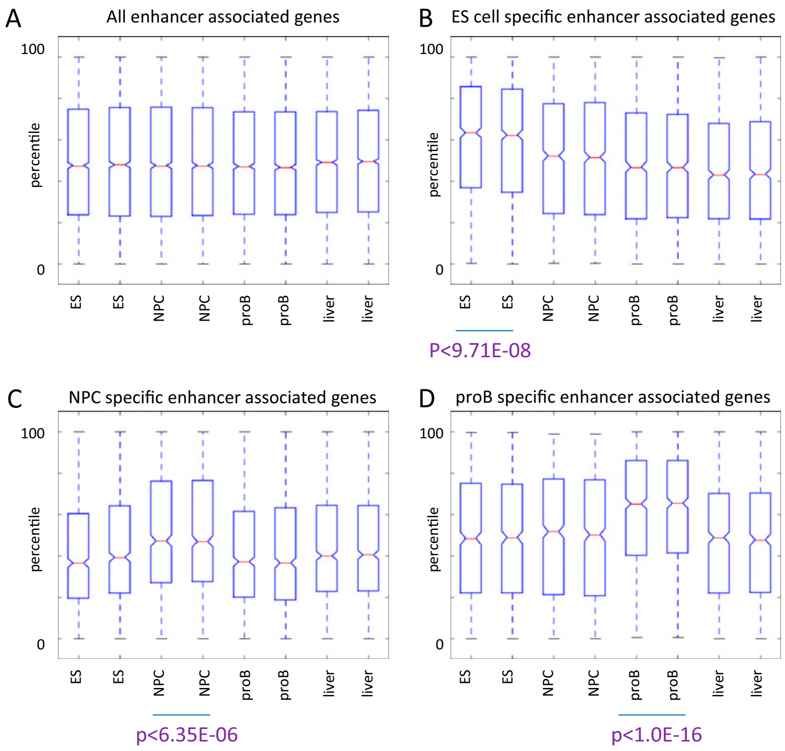

Fig. S4. Genes that are enhancer proximal in a cell type-specific manner are selectively more active. (*A*–*D*) Gene expression microarray data displayed in box-plots displayed for select genes found specifically associated with enhancers in the indicated cell types only. The different boxes display how these genes behave in the cell types tested in duplicate as shown along the axis. Solid boxes display the 25–75% of ranked genes, with the mean indicated as an intersection. *P* values are from Bonferroni-corrected Mann–Whitney *U* tests for the underlined samples being more active than any of the other samples.

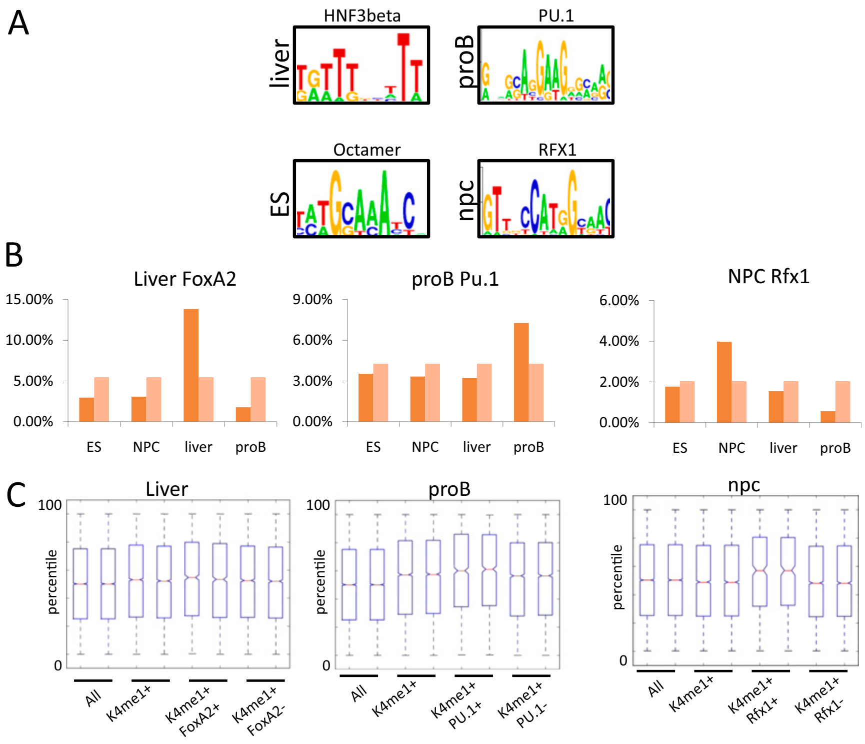

Fig. S5. Proximal genes are not consistently affected by diverse cell-specific transcription factors found at enhancers. (*A*) Motif search analysis using conserved sequence under the K4me1 enriched regions in the cell types indicated. (*B*) Enrichment graphs for the indicated transcription factors in the four cell types indicated (*x* axis). *y* axis displays percentage of cell-specific enhancers. Light bars show percentage enriched expected by chance, dark bars show actual percentage enriched. (*C*) Gene expression of microarray data generated in duplicate from the different tissues displayed on top of the graph. Box-plots show all genes (All) and genes found specifically associated with enhancers in liver (H3K4me1) either enriched (+) or unenriched (–) for the indicated transcription factors. Solid bars of boxes display the 25–75% of ranked genes, with the mean as an intersection. *P* values in the text are from Bonferroni-corrected Mann–Whitney *U* tests.

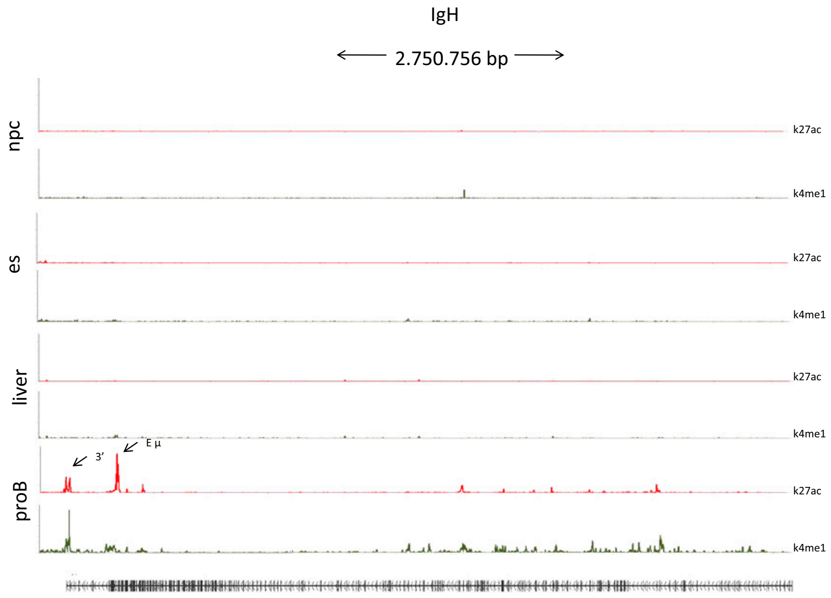

Fig. S6. Specific marking of known active enhancer regions in the Ig heavy chain locus. Browser track for H3K4me1 (green) and k27ac (red) showing the IgH locus (known active enhancers are indicated by arrows) for the indicated cell types (scale 1–100 for H3H3K4me1 and 1–200 for H3K27ac).

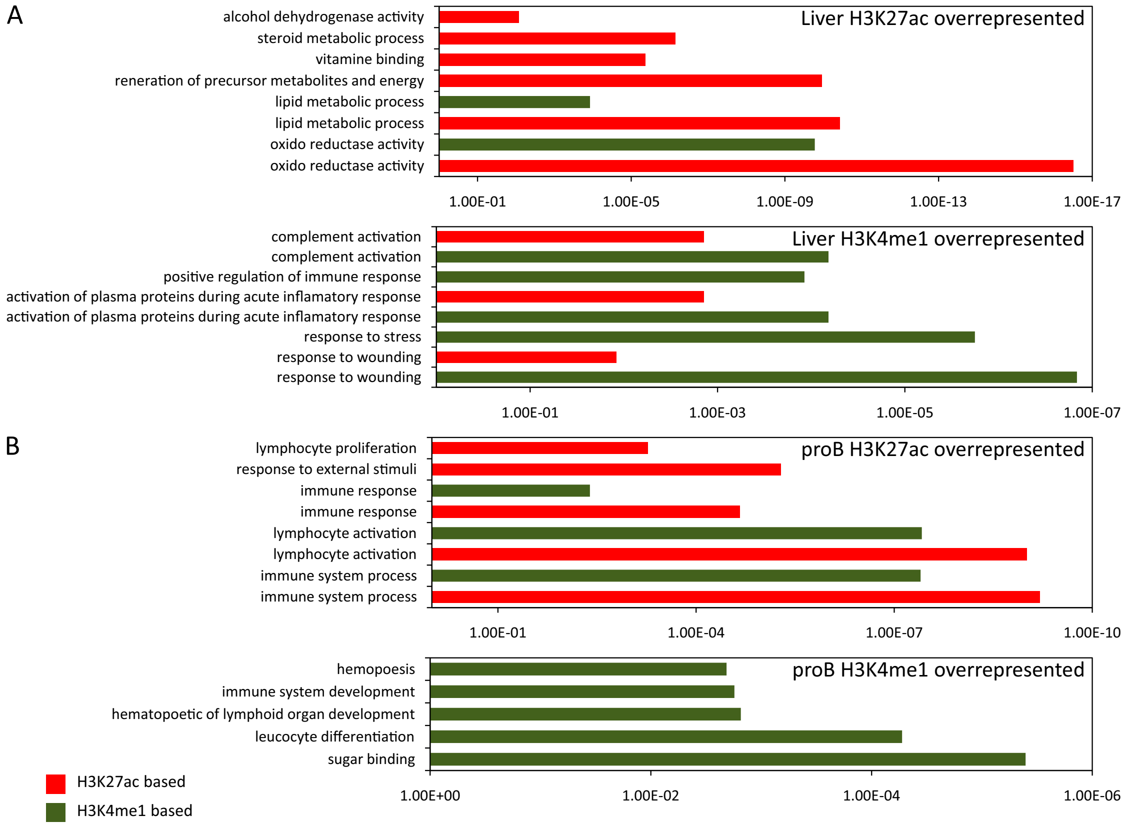

Fig. S7. Active enhancers correlate differently to gene functions compared with inactive enhancers. (*A* and *B*) GO analysis displaying gene functions for genes specifically associated with enhancers in the indicated cell types in (*A*) liver and (*B*) proB. Red bars display functions based on proximal genes to H3K27ac-positive enhancers. Green bars display functions based on proximal genes to H3K4me1-positive enhancers. *P* values are displayed on the *x* axis. Functions for each cell type are divided in two graphs. H3K27ac overrepresented (*Upper*) shows functions associated specifically or with higher confidence to genes proximal to H3K27ac-enriched enhancers. H3K4me1 overrepresented (*Lower*) shows functions associated specifically or with higher confidence to genes proximal to H3K4me1-enriched enhancers.

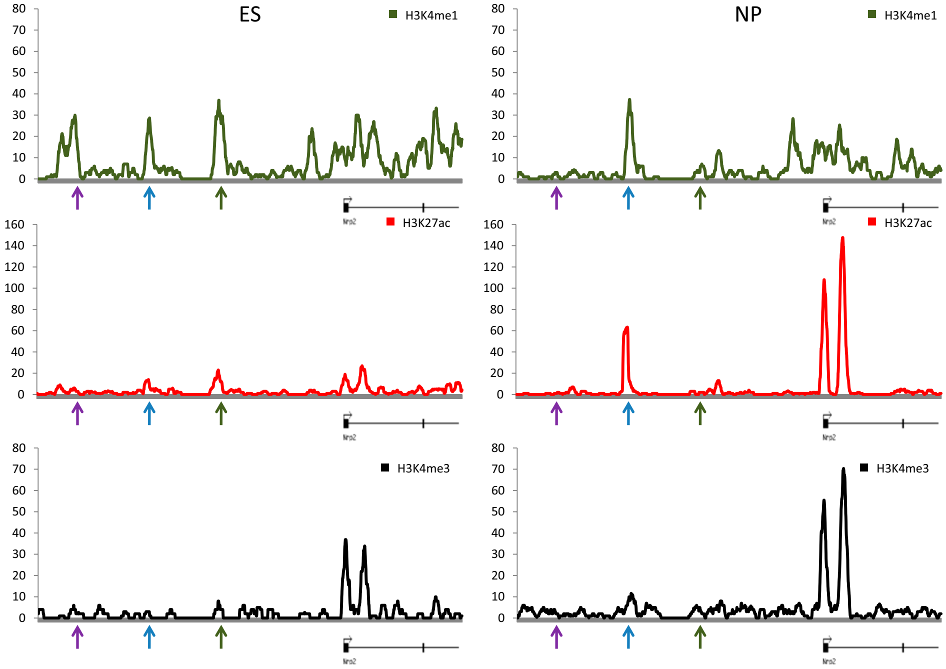

Fig. S8. Neurophilin-2 is associated with poised enhancers in ES cells that become active in neural progenitors (NPs). Gene track for H3K4me1 (green, *Top*), H3K27ac (red, *Middle*), and H3K4me3 (black, *Bottom*) for a region containing the Neurophilin-2 gene. Tracks are shown in ES cells (*Left*) and neural progenitor cells (*Right*). Arrows indicate poised enhancers in ES cells, one of which gains the H3K27ac mark in NPs, whereas the other two are decommissioned. *y* axis shows number of reads.

Table S1. H3K4me1- and H3K27ac-enriched promoter distal elements and proximal genes in different tissues, Rfx1-enriched regions in neural progenitor cells, H3K4me1-enriched regions in induced pluripotent stem cells, and p300-enriched regions in ES cells

Table S1