Supplemental materials

Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression

Di Huang[1], Hanna M Petrykowska[2], Brendan F. Miller [2], Laura Elnitski[2]*, Ivan Ovcharenko[1]*

Histone modification ChIP-seq and RNA-seq data across cells/tissues

We downloaded the epigenetic and RNA-seq data from the Epigenomics roadmap project:

- http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak;
- http://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz

The H3K27me3-DHSs were defined as the DNase-seq peaks overlapping with the H3K27me3 ChIP-seq peaks. The requirements of DNase-seq, RNA-seq, and H3K27me3 ChIP-seq data restricted us to 27 tissues. We then filtered out the H3K27me3-DHSs with the length of <200 bp, and those of which the centers are located within gene promoters. After collecting H3K27me3-DHSs across these tissues, we merged the overlapped or immediately-neighboring (the gaps between genomic regions < 100bp) H3K27me3-DHSs using "bedtools merge" with the setting of n = 100. We next measured the similarity of H3K27me3-DHS maps across the 27 tissues and discarded fetal small intestine and variant human mammary epithelial cell (vHMEC) since they were greatly redundant to fetal intestine large and HMEC, respectively. We ended up with epigenetic and transcriptional data for 25 distinct cell lines (Fig. S2).

TFBS profiles based on TF ChIP-seq data

We downloaded TF ChIP-seq data from the ENCODE project (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/), and focused the human lymphoblastoid cell line GM12878, the human liver cancer cell line HepG2, and the human cervical cancer cell line HeLa S3 cell line due to the abundance of data for these cell lines. For a TF having multiple ChIP-seq datasets from different labs or conditions, we collected all the ChIP-seq peaks to establish a unified TFBS profile.

We used posCORs as a background to evaluate the TFBS signature of silencers. To ensure reliability of analysis, we focused only on the TFs for which TFBSs occurred in at least 3% of either enhancer sequences or silencer sequences. For example, in GM12878, P300 was excluded from our study since less than 3% of H3K27me3-DHSs overlap with P300 ChIP-seq peaks.

Bivalent elements defined by chromHMM

Bivalent elements, carrying the activating and repressive histone modifications simultaneously, were defined by Chromatin Hidden Markov Model (ChromHMM). Through exploring histone modification profiles, ChromHMM annotated human genomic regions with 15 functional categories, including enhancers, transcriptional start sites, bivalent regions, etc. Bivalent regions, i.e., the functional category featuring the co-existence of activating and repressive histone modifications, was used to verify the transcriptional impact of the predicted silencers. Since promoter regions were not the focus, only the DNA regions labeled as bivalent enhancers were used in this study. We demonstrate that negCORs and bivalent enhancers have disparate transcriptional impacts.

Hi-C data

Hi-C data that we used were deposited to Gene Expression Omnibus (GEO GSE63525) by Aiden's lab. We downloaded Hi-C loops in seven human cell lines, i.e., GM12878, mammary epithelial cells HMEC, umbilical vein endothelial cells HUVEC, HeLa S3, myelogenous leukemia cells K562, normal epidermal keratinocytes NHEK, and fetal lung cells IMR90. Whereas the first six cell lines were examined in our study, IMR90 was used as the surrogate of normal human lung fibroblasts NHLF due to the close match of these two cell lines. The reported Hi-C loops were used to verify our prediction results in the corresponding cell lines. A silencer, or more broadly a regulatory element (RE), was considered to target a gene when the center of that silencer and the transcription start site (TSS) of the inquiry gene residing in two ends of a Hi-C loop.

Hi-C loops were also used to analyze the collective effect of REs on gene expression. Given a gene, we built its RE compound by collecting all the silencers and enhancers connected to it by Hi-C loops. For an RE compound, the enrichment of the silencers (enhancers) was evaluated with eq. 1 (eq.2) by setting $s(e)$ the numbers of the enhancers (silencers) included in the inquiry RE compound, with $S(E)$ being the total numbers of the silencers (enhancers) having Hi-C connection. The limited availability of Hi-C loops caused relatively low enrichment estimates. As such, a loose criterion was used to determine if an RE compound was silencer/enhancer-rich. That is, an RE compound was labeled as silencer-rich when $P_r^S(X > e) < 10^{-2}$ and $s \geq 1$.

GTEx eQTLs

Genotype-Tissue Expression project (GTEx) release V6p was explored in this study. GTEx eQTL studies associate noncoding SNPs with the genes when the genotypic changes of SNPs significantly correlate with the transcriptional variations of genes, suggesting the cis-/trans-regulatory relationship between the SNPs and the associated genes. In total, seven of our cell lines were identical to or matched with one GTEx tissue each. HepG2 in our study matched with liver in GTEx, pancreas with pancreas, small intestine with small intestine, GM12878 with whole blood, NHLF with lung, psoas muscle with muscle, and human skeletal muscle myoblast HSMM with muscle.

Given a SNP-gene association, we linked the silencers (and other REs) hosting the SNP with the gene. These genes were used to evaluate the impact of predicted silencers (and other REs) in the same way of Hi-C target genes for evaluating the regulatory function of negCORs and SVM silencer predicts.

Performance evaluation of SVM classification

To test classification performance of SVMs, a five-fold cross validation scheme was used. In this scheme, a training dataset was equally divided into five subsets. After using every subset for validating the SVM built on the other four subsets, we obtained validation results on all training samples and evaluated these results in terms of false positive rate, precision, and recall. Also, we applied all the built SVMs to score each of the H3K27me3-DHSs other than negCORs and posCORs and used the average of the SVM scores as a final estimate of the inquiry H3K27me3-DHS. H3K27me3-DHSs having scores greater than a threshold were then marked as potential silencers. The threshold was determined in such a way that the false positive rate on the validation results was 0.1. After concatenating negCOR and SVM silencers, we delivered a silencer map for each tested cell line.

Sharpr-MPRAs

We explored the results of Sharpr-MPRAs (Massively Parallel Report Assays) to examine the function of the REs of interest. In Sharpr-MPRAs, each of 15,720 DNA fragments (defined by the DNase-ChIP peaks in four cell lines) was coupled with a promoter and a distinct reporter gene. The activity of a tested DNA sequence was scored based on the normalized expression of the corresponding reporter gene. Following the rules used by the Sharpr-MPRA designers, we calculated the MaxPos for each Sharpr-MPRA sequence in K562. A negative MaxPos score implies a regulatory repressive impact. After overlaying the Sharpa-MPRA sequences with the predicted silencers in K562, we retained the ones having the >80% overlap with a predicted K562 silencer. The MaxPos scores of these Sharpa-MPRA sequences provided the assessment of K562 silencers. We repeated the above process on the H3K27me3-DHSs, H3K27ac peaks and DNase-ChIP peaks in K562. The comparative results demonstrated the function of RE groups.

Human TFs

We collected human TFs and transcription co-factors (TCF) from the Dragon Database for Human Transcription Co-Factors and Transcription Factor Interacting Proteins (TcoF-DB) resource.

GWAS SNPs

We downloaded the NHGRI GWAS Catalog of April 2015, where 14,841 GWAS SNPs were associated with at least one of 1,106 traits. To account for the incompleteness of the SNPs directly assayed in GWAS studies, we expanded the SNP set through identifying all SNPs in a tight link disequilibrium block with each GWAS SNP ($r^2 > 0.8$ and $distance < 500\ kb$) based on at least one population from the 1000 Genomes Project , such as Northern Europeans from Utah (CEU), Yoruba in Idaban, Nigeria (YRI), and Han Chinese in Beijing, China/Japanese in Tokyo, Japan (CHB/JPN) by using Single Nucleotide Polymorphism (SNP) Annotation and Proxy Search (SNAP) (Johnson et al. 2008). After that, we linked these tight-LD SNPs to the corresponding traits and eventually obtained 324,454 SNPs associated with 1,106 GWAS traits.

In total, 19,937 of the GWAS SNPs reside in silencers. Given the extreme diversity of GWAS traits and the relative scarcity of GWAS SNPs, we used all the GWAS SNPs to evaluate the silencers as a whole, instead of performing analyses in individual cell lines.

Mapping predicted silencers from human genome assembly hg19 to GRCh38

The prediction of negCOR and SVM silencers are based on human genome assembly hg19 due to significant data abundancy on hg19 as compared to on GRCh38. We will lose about 0.4%

predicted silencers after aligning them to the newest version of human genome (i.e. GRCh38), which would not change our conclusion.

Supplemental Figures

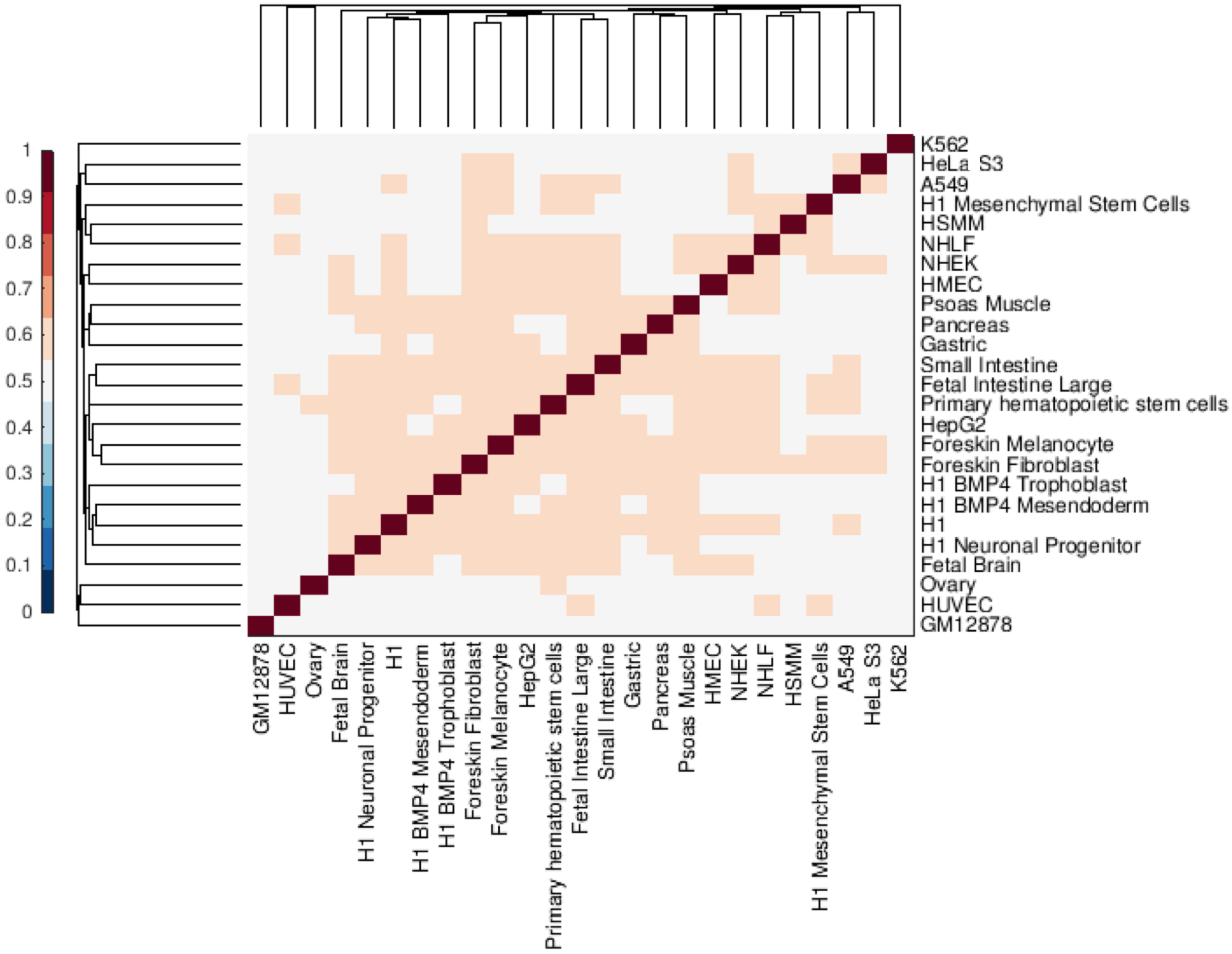

Figure S1. Similarity of H3K27me3-DHS activity profiles across 25 cell lines.

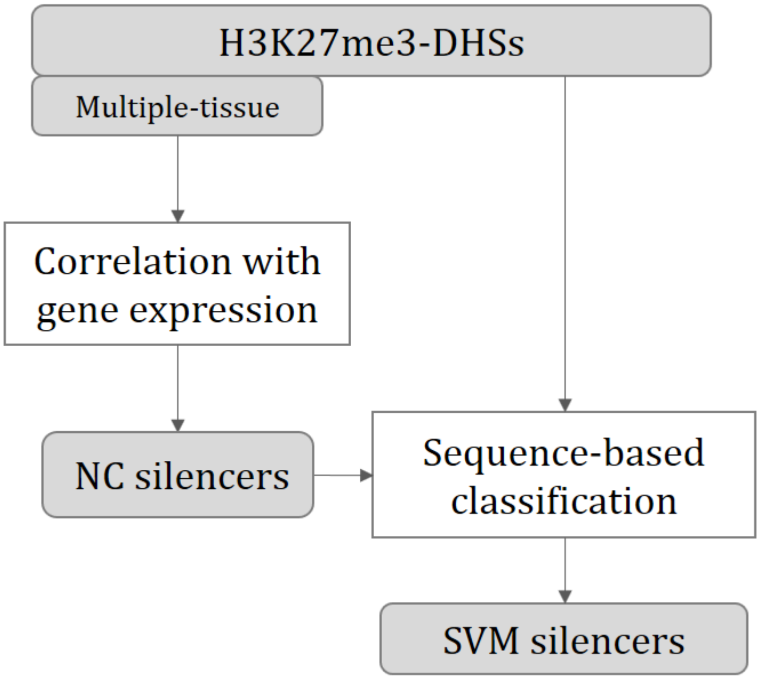
H3K27me3-DHSs
Multiple-tissue
Correlation with gene expression
NC silencers
Sequence-based classification
SVM silencers

Figure S2. Schematic overview of the proposed silencer identification framework.

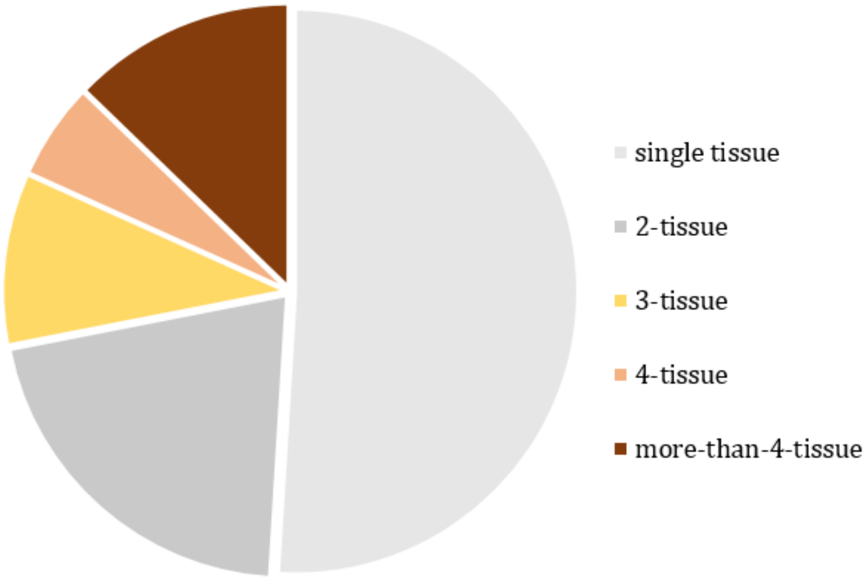

Figure S3. Distribution of tissue specificity of H3K27me3-DHSs.

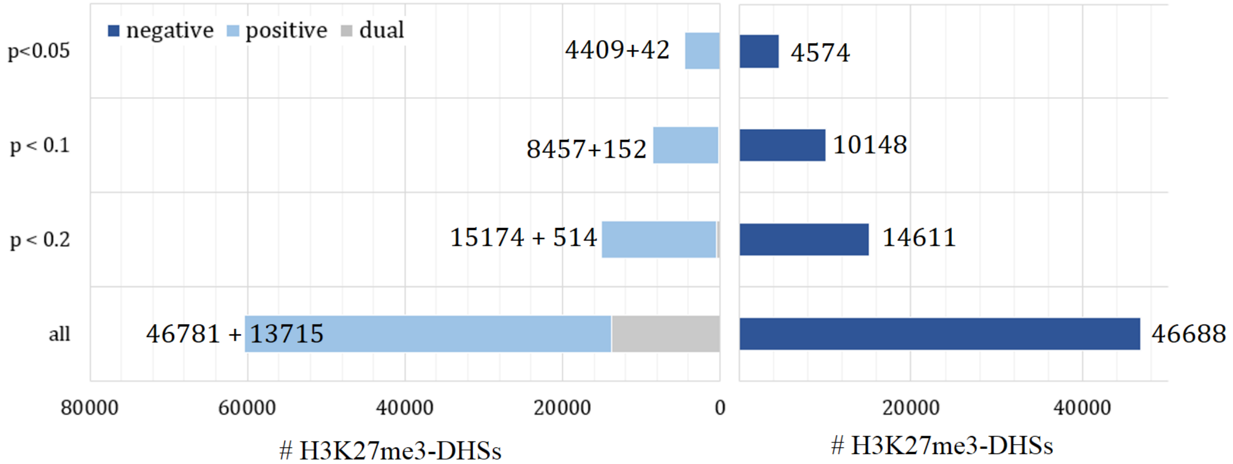

Figure S4. Distribution of the correlation between the activity of H3K27me3-DHSs and the expression of their proximal genes. The significance p values were estimated under the Wilcoxon rank-sum test. An intronic H3K27me3-DHS was associated with its host gene, while an intergenic one was assigned with two genes, one in each genomic direction (downstream and upstream). "Dual" H3K27me3-DHSs are those negatively correlated with one assigned gene, and positively correlated with the other. $p < 0.05$ was used to identify the significantly negatively correlated H3K27me3-DHSs (negCORs) and significantly positively correlated H3K27me3-DHSs (posCORs). To avoid ambiguity, we conservatively excluded dual H3K27me3-DHSs from the sets of negCORs and posCORs.

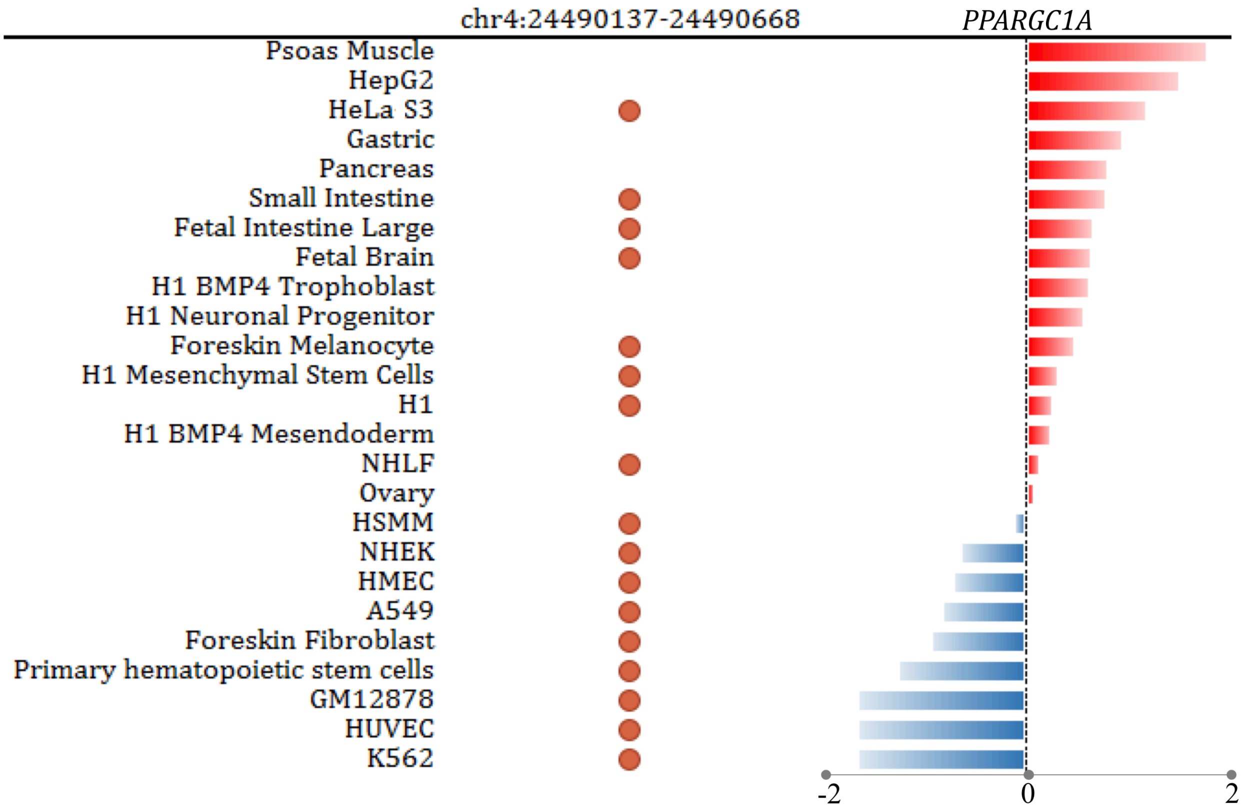

Figure S5. A negCOR silencer in the neighborhood of the *PPARGC1A* gene.

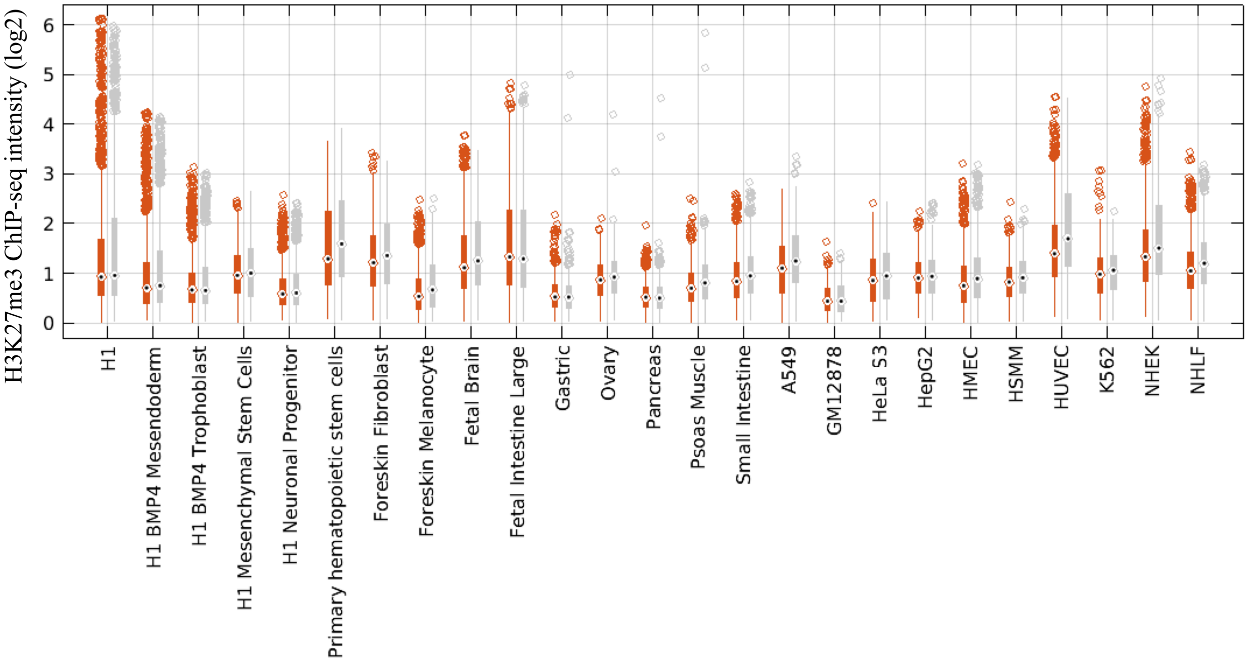

Figure S6. Signal intensity of ChIP-seq H3K27me3 of negCORs (in orange) and posCORs (in grey) across cell lines. The median and standard deviations of signal intensities are represented by the dot-center diamonds and the bars flanking the diamonds, respectively.

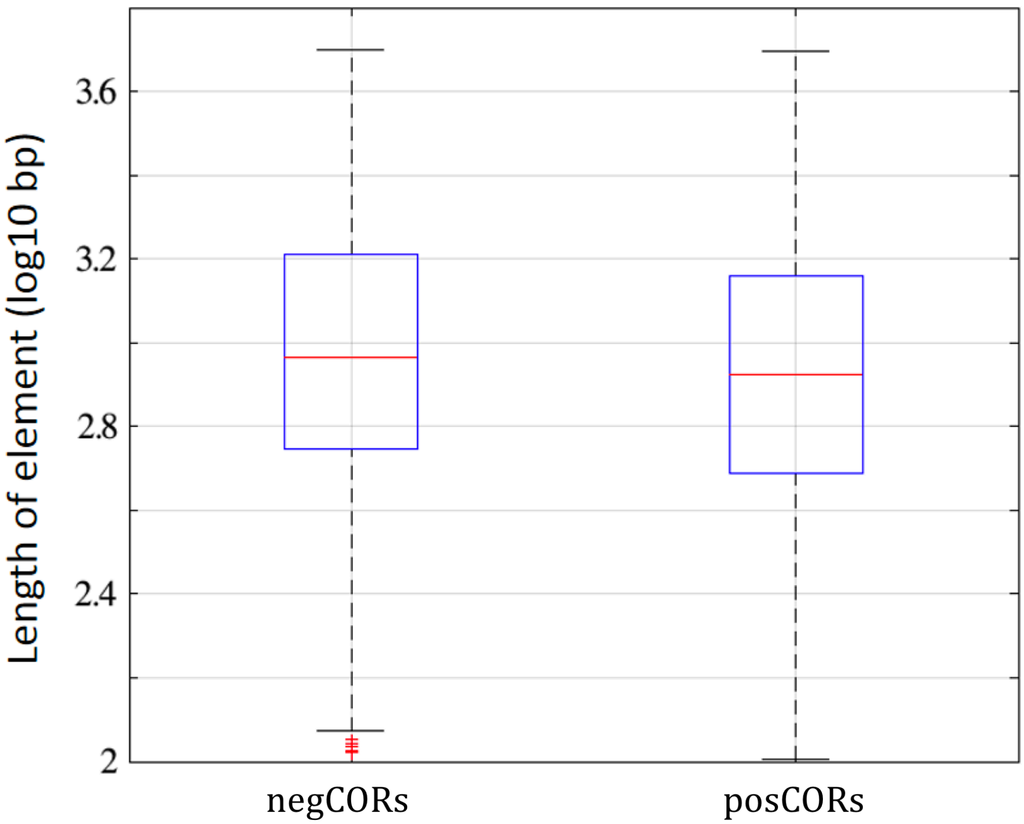

Figure S7. Lengths of negCORs and posCORs.

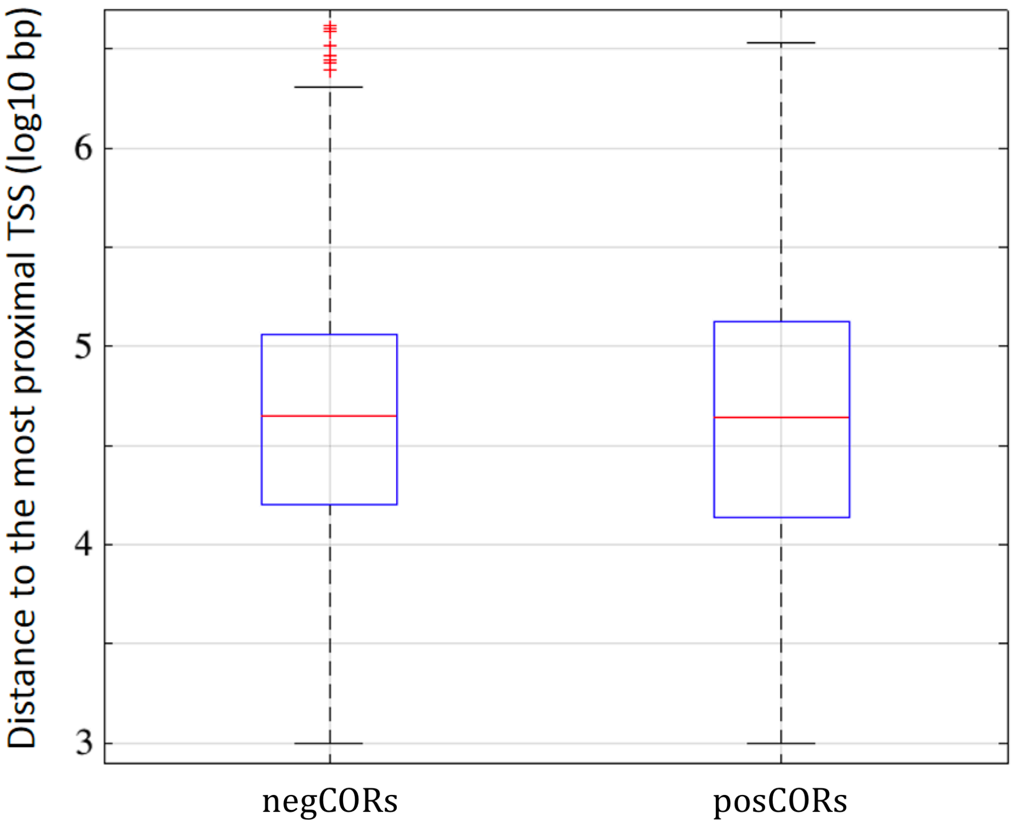

Figure S8. Distances of H3K27me3-DHSs to their nearest transcriptional start sites (TSSs).

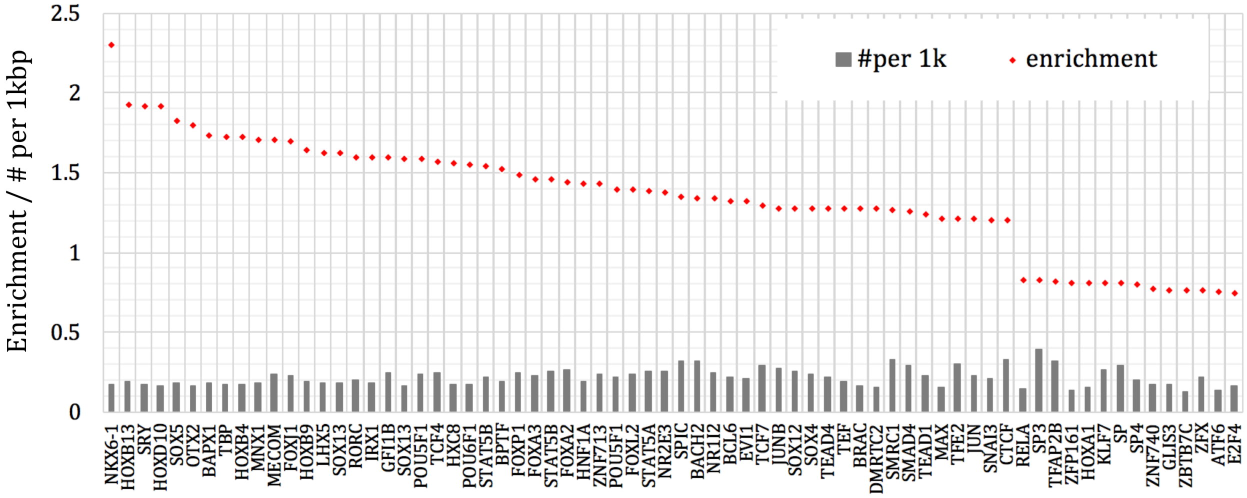

Figure S9. Motif-based TFBS enrichment and density (# per 1kbp) of negCORs in K562 cell line.

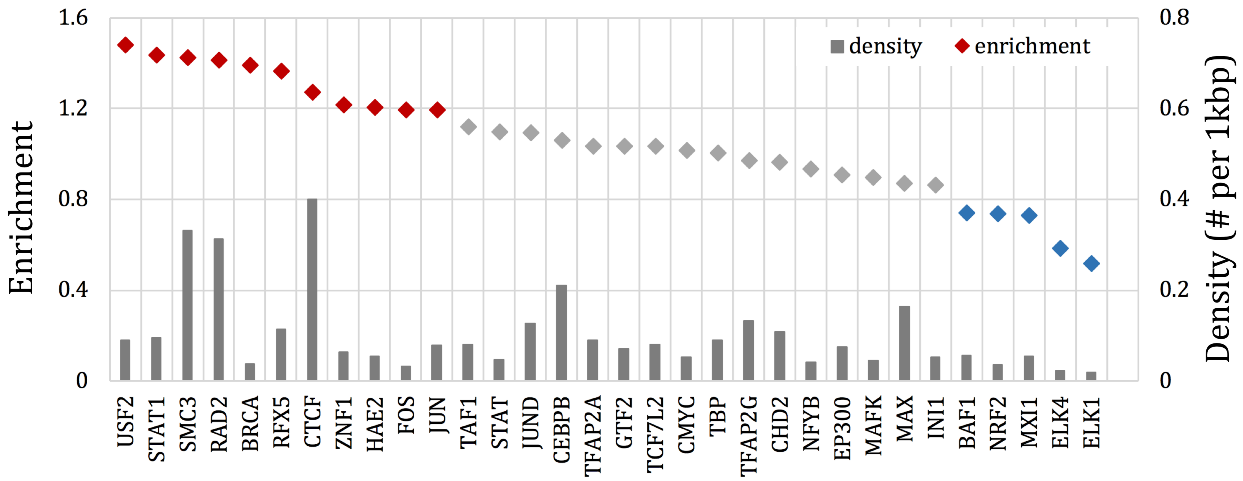

Figure S10. Enrichment of ChIP-seq TFBS signatures of negCORs in HeLa-S3.

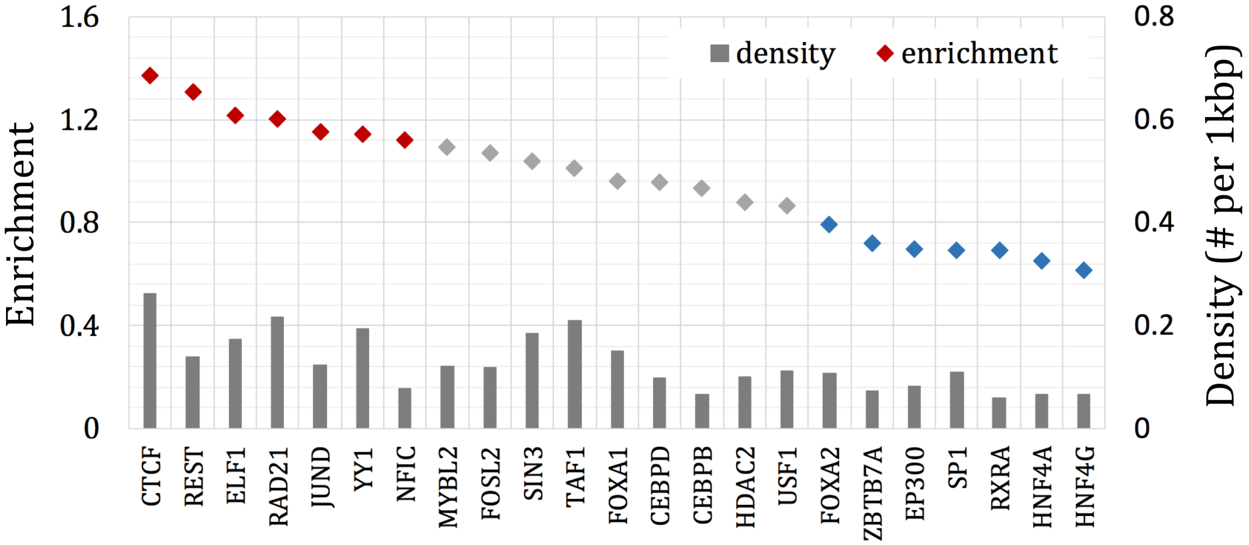

Figure S11. Enrichment of ChIP-seq TFBS signatures of negCORs in HepG2.

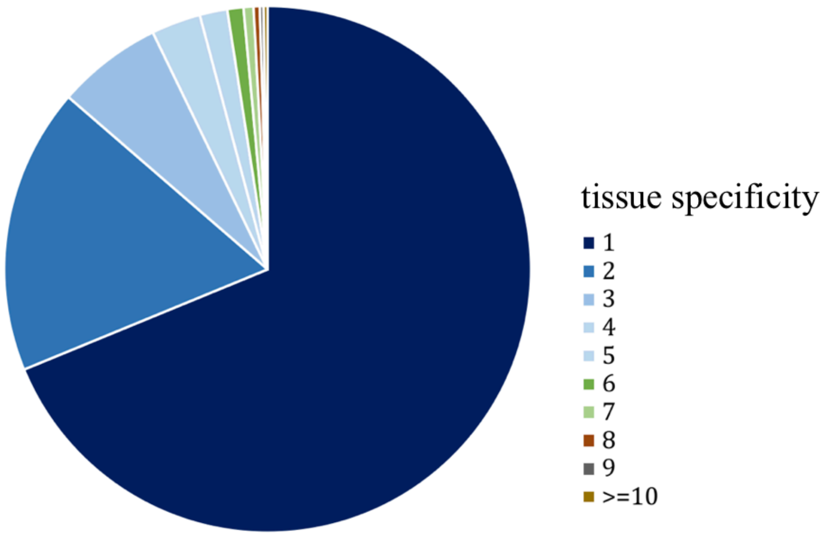

Figure S12. Tissue specificity of SVM silencers. Tissue specificity of an SVM silencer is measured as the number of the cell lines in which the inquiry silencer is present.

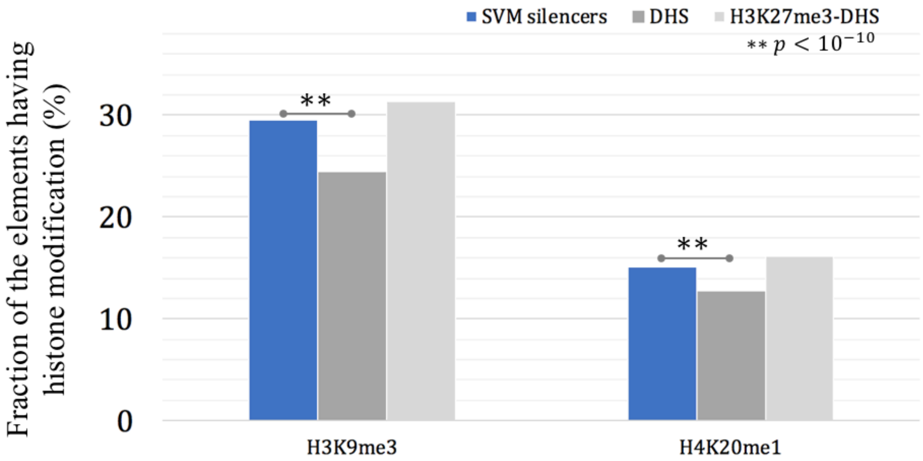

Figure S13. Density of H3K9me3 and H4K20me1, the repression-associated histone modifications, along SVM silencers, DHSs and H3K27me3-DHSs. DHSs/H3K27me3-DHSs used here were randomly selected from all the DHSs/H3K27me3-DHSs active in the corresponding cell line and in the matching length to inquiry SVM silencers.

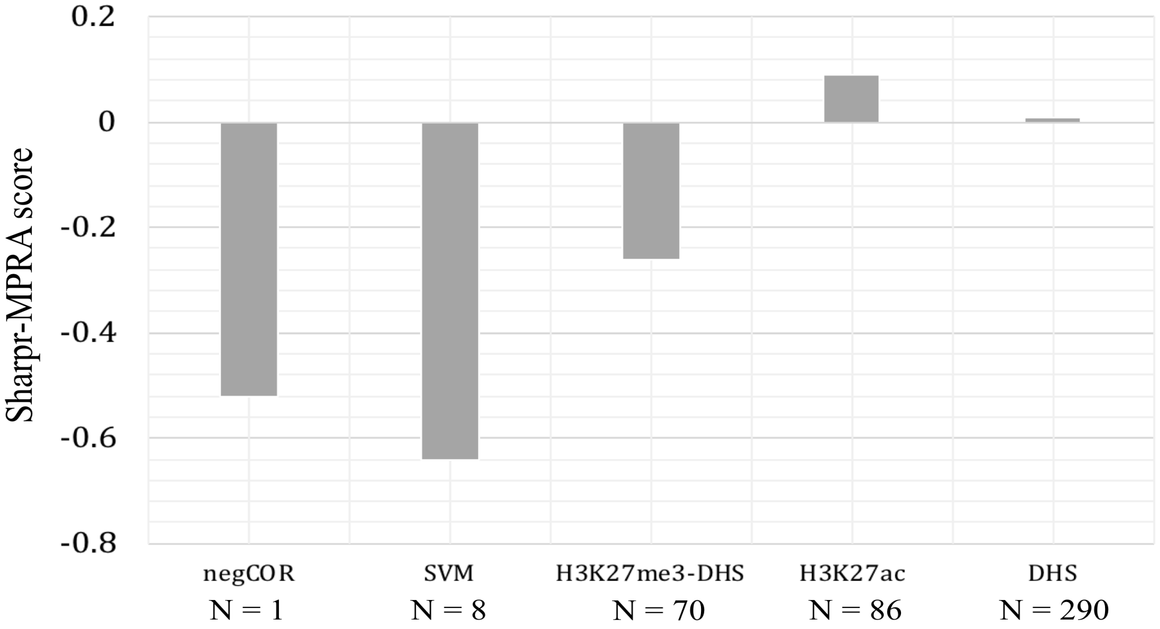

Figure S14. Sharpr-MPRA score of predicted silencers in K562 cell line. “N=*” gives the number of the elements examined in Sharpr-MPRAs.

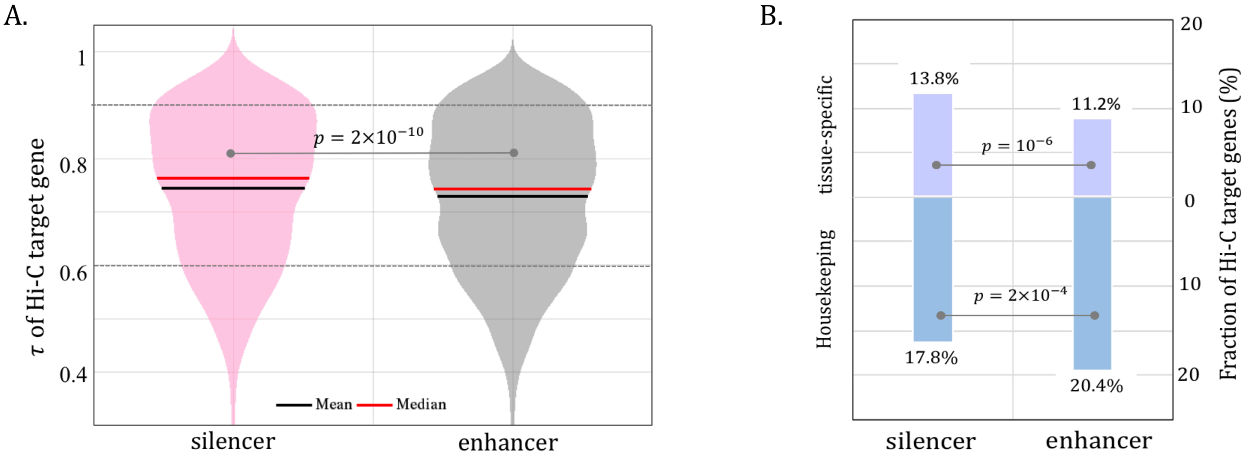

Figure S15. Tissue-specificity of genes linked to silencers and enhancers by Hi-C connections. τ of genes targeted by silencers and enhancers (A). Fraction of tissue-specific ($\tau > 0.9$) and housekeeping ($\tau < 0.6$) genes having Hi-C links to silencers and enhancers (B).

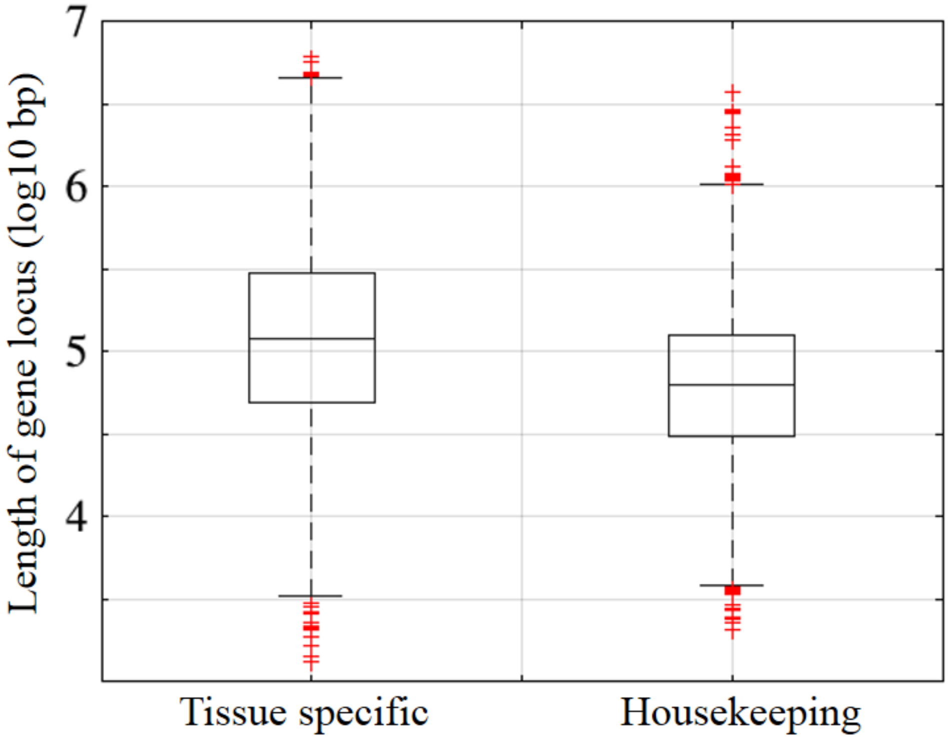

Figure S16. Length of tissue-specific or housekeeping gene locus.

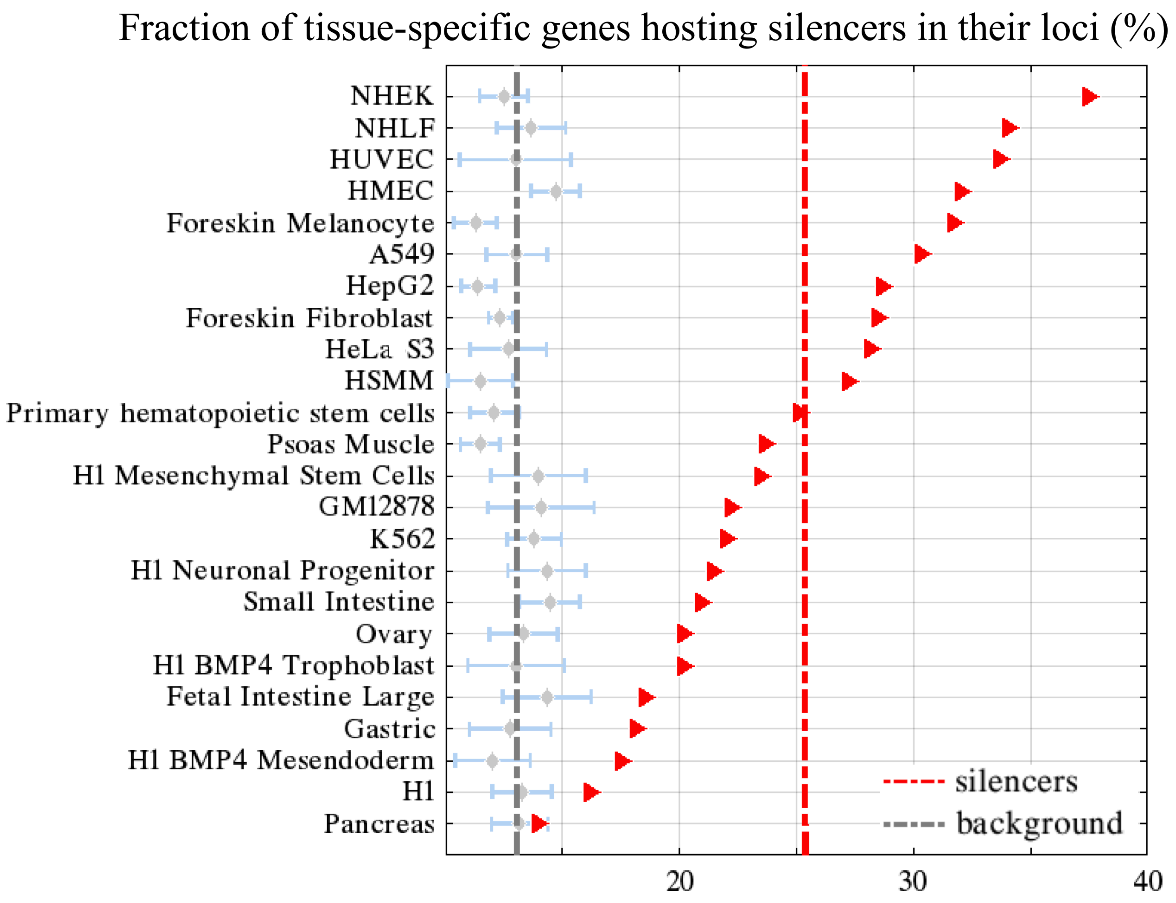

Figure S17. Fraction of tissue-specific genes hosting silencers in their loci. Background was generated through randomly selecting the H3K27me3-DHSs with the matching numbers to the corresponding silencers. The grey diamonds and the flanking blue lines are the median and standard deviations of the results on 50 independent background sets, respectively.

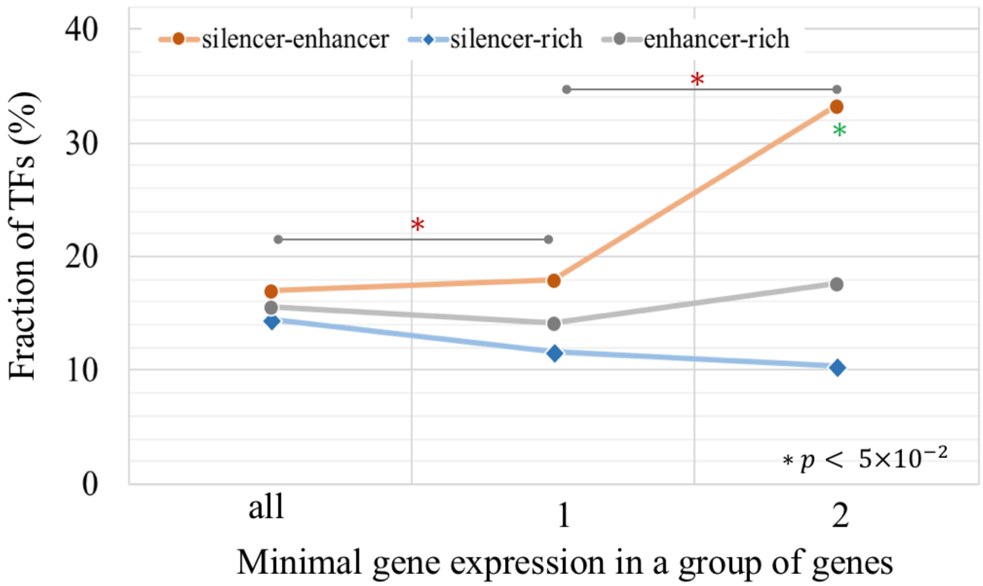

Figure S18. Fraction of TFs among different genes stratified based on their composition of the silencers and the enhancers linked to them by Hi-C connections. Red asterisks indicate the significant differences of TF fraction in silencer-enhancer-loci across gene groups, while green asterisk suggests the significant difference between silencer-enhancer-loci and silencer-rich loci.

Supplemental Tables

Table S1. Basic characteristics of negCORs.

	Repeat density (%)	GC content (%)	intronic RE (%)	CpG island (%)
negCOR	26.4	49.7	36.8	7.87
posCOR	27	52.1	31.8	7.76
H3K27me3-DHS	31.2	47.4	46.9	2.58
enhancer	33.5	45.6	53.1	2.28

Table S2. NegCORs silencers and silencers predicted using SVMs. Each row corresponds to a silencer. Three columns are “cell name”, “silencer coordinates” and “prediction method”, respectively. **The table is provided as a separate text file Supplemental_Table_S2.txt.**

Table S3. Fraction of REs carrying polycomb associated ChIP-seq peaks (%). Here, the enhancers are identified as the DHSs carrying H3K27ac and H3K4me1 modification.

REs	H3K27me3	EZH2	SUZ12
negCOR silencers	100	57.8	13.3
SVM silencers	100	53.9	13.2
H3K27me4-DHS	100	54.5	11.1
Enhancers	25.6	37.6	5.7
DHSs without H3K27me3	0	28.6	5.1

Table S4. Experimental results

element	coordinate	Length(bp)	nearest gene	R1	R2	R3	Avg	std
S1	chr12:66377281-66377852	572	HMGA2 (chr12:66,219,051-66,357,072; +)	1.339	1.395	1.417	1.384	0.032838
S2	chr7:54856104-54856781	677	EGFR (chr7:55086714-55275773; +)	1.625	1.699	1.818	1.714	0.079503
S3	chr1:27293222-27293922	700	TRNP1 (chr1:27320195-27327377; +)	1.611	1.253	1.428	1.431	0.146165
S4	chr1:178470782-178471226	444	RASAL2 (chr1:178,310,731-178,442,374; +)	1.7	1.75	1.93	1.79	0.098826
S5	chr9:91692203-91692813	610	SHC3 (chr9:91,628,362-91,793,375; -)	1.577	0.962	1.721	1.42	0.329147
S6	chr7:100787036-100787976	940	SERPINE1 (chr7:100,770,370-100,782,547; +)	1.759	1.723	1.728	1.736	0.015937
S7	chr1:228122078-228123263	1185	WNT9A (chr1:228106357-228135631; -)	0.906	0.862	0.8	0.856	0.043482
S8	chr15:52773630-52774898	1268	MYO5A (chr15:52599480-52821247; -)	2.399	2.279	2.314	2.331	0.050388
S9	chr1:120533133-120533497	364	NOTCH2 (chr1:120,454,176-120,612,317; -)	1.272	1.239	1.406	1.306	0.072215
S10	chr1:207084179-207084844	665	FAIM3 (chr1:207,078,364-207,087,30; -)	0.493	0.589	0.592	0.558	0.045978
H1	chr1:222789849-222790431	582	TAF1A (chr1:222731244-222763275; -)	1.75	1.708	1.633	1.697	0.048394
H2	chr10:72003325-72003761	436	PPA1 (chr10:71,962,586-71,993,190; -)	1.924	2.047	1.985	1.985	0.050216
H3	chr12:4737219-4738051	832	AKAP3 (chr12:4,724,676-4,754,358; -)	1.58	1.587	1.616	1.594	0.015588
H4	chr21:35068167-35068785	618	ITSN1 (chr21:35,014,784-35,210,802; +)	2.374	2.601	2.531	2.502	0.094914

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**: 2938-2939.