Enhancer-silencer transitions in the human genome

Di Huang and Ivan Ovcharenko*

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA.

* To whom correspondence should be addressed: ovcharen@nih.gov

Running title

Enhancer-silencer transitions

Table of Contents

Supplementary Notes

CNN model

The CNN model (Fig. 1A) consists of five convolutional layers and two fully-connected layers arranged sequentially. The details of this model are:

1. 1-dimensional (1D) convolutional layer with 480 kernels, each having a window size of nine and a step size of one.
2. Maxpooling layer with a window size of nine and a step size of three.
3. Dropout layer with a dropout proportion of 0.2.
4. 1D convolutional layer 480 kernels, each having a window size of four and a step size of one.
5. Maxpooling layer with a window size of four and a step size of two.
6. Dropout layer with a dropout proportion of 0.2.
7. 1D convolutional layer with 240 kernels, each having a window size of four and a step size of one.
8. Maxpooling layer with a window size of four and a step size of three.
9. Dropout layer with a dropout proportion of 0.2.
10. 1D convolutional layer with 320 kernels, each having a window size of four and a step size of one.
11. Maxpooling layer with a window size of four and a step size of three.
12. Dropout layer with the dropout proportion of 0.2.
13. 1D convolutional layer with 320 kernels, each having a window size of four and a step size of one.
14. Maxpooling layer with a window size of four and a step size of three.
15. Fully connected layer of 180 neurons with the sigmoid activation function.
16. Fully connected output layer of 3 neurons with the SoftMax activation function.

We used the Rectified Linear Unit (ReLU) activation function in the convolutional layers. In the convolutional and fully connected layers, the penalty coefficients of L1 and L2 regularizations were 10^{-8} and 5×10^{-8}, respectively, and the max weight constraint of the parameters in a kernel or neuron was 0.9.

On average, our training datasets consist of 6% enhancers, 9% silencers, and 85% background samples across the tested cell types. Imbalance among sample classes is a naturally inherent issue in genome-wide predictions of silencers/enhancers as only a small fraction of human noncoding DNA consists of gene regulatory elements (Singh et al. 2018). Although the class imbalance ratio in training datasets is moderate for large-data Deep Learning (Johnson and Khoshgoftaar 2019), we evaluated the possible influence of class imbalance on the reported results. The original cost function, in which all samples were penalized equally, was compared with the weighted-class cost function, in which heavy penalty parameters were assigned to minor-class samples (i.e., enhancer and silencer samples). We utilized the function "class_weight" from the Python library scikit-learn (Pedregosa et al. 2011) to estimate class weights. Across the tested cell types, the weighted-class CNN models demonstrate similar performance on test samples as the corresponding original CNN models, with the differences in AUC ROC and AUC PRC scores having the average of -0.0004 and the standard deviation of 0.039 (Fig. S18A). On the experimentally validated K562 silencer sets, the original model CNN delivers a slightly better performance than the weighted-class CNN model (Fig. S18B). Furthermore, on average, 87% of enhancers and 69% of silencers predicted by the original CNN models were also labelled as the corresponding class by the weighted-class CNNs (using the output of FDR = 0.1 on test samples as prediction cutoffs). Finally, the weighted-class CNN-SASs of raQTLs are highly correlated with the original CNN-SASs ($r = 0.81, p = 0$, Fig. S18C). To sum up, the presented results consistently demonstrate a strong match in the output between the weighted-class and original CNN models, with a slightly better performance of the original CNN model on the experiment-validated silencers. Therefore, the original CNN models were built and utilized by this study, in a manner similar to previously published studies (Zhou and Troyanskaya 2015; Kim et al. 2016).

Data for training CNN models

We downloaded DNase-seq peaks (1) from the Roadmap Epigenomics Project (http://egg2.wustl.edu/roadmap) and unified all peaks of 1,000 bp ($midpoint \pm 500\ bp$). To define the function of these DNase-seq peaks, we overlapped them with ChIP-seq peaks of histone marks, including H3K27ac, H3K27me3, H3K4me1, and H3K4me3. A DNase-seq peak was considered to carry a histone mark when its central section ($midpoint \pm 200\ bp$)

overlapped with the ChIP-seq peaks reported for the same cell type. The DNase-seq peaks containing H3K27ac but no H3K27me3 signals were considered as enhancer candidates. The silencer training samples came from two sources: 1) the DNase-seq peaks carrying H3K27me3 but neither H3K27ac nor H3K4me1/3 peaks; 2) the H3K27me3 ChIP-seq peaks carrying no DNase-seq or H3K27ac or H3K4me1/3 signals. H3K27me3 ChIP-seq peaks were also extended into the lengths of 1,000 bp. The primary T cells and embryonic stem cells studied here are the cell types E034 and E003 in the Roadmap Epigenomics Project, respectively.

Gene expression data and gene annotations

We downloaded gene expression data from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium 2015) at http://egg2.wustl.edu/roadmap/data/byDataType/rna/ expression. Gene expression, measured as the Reads Per Kilobase of transcript per Million mapped reads (RPKM), was normalized so that the expression level genes had a median of zero and a standard deviation of one across cell lines. A positive/negative normalized level of expression was thus indicative of a gene being highly or lowly expressed in the corresponding cell line. The gene annotations we used were downloaded from the GENCODE project (Frankish et al. 2019).

Genomic mappability

We downloaded genomic mappability scores (GMSs) from the ENCODE project (wgEncodeDukeMapabilityUniqueness20bp.bigWig) (Boyle et al. 2008). Bases having GMSs of 1 were considered as certainly mapped. Given a sequence, the fraction of certainly mappable bases was calculated. A sequence was considered as high GMS when its GMS fraction was greater than 50%.

CNN-based silencing odds ratio of mutations

We first derived the probability function of $(ys - ye)$ using test samples. Here, ys and ye are the silencing and activating capability of a given sequence, respectively. With the Fitter Python library (https://github.com/cokelaer/fitter/pull/37), we adjusted different univariate distributions (including normal, exponential, T, gamma, beta, log-normal, double Weibull, generalized extreme value, and Pareto distribution) to fit the distribution of $(ys - ye)$ values. A T distribution function is the best fit to the distribution of $(ys - ye)$ values (Fig. S12A where only top-5 best fitting functions are shown). With the derived function, we then evaluated the odds ratio of silencing capability of a sequence x as

$$OR\_silencing(x) = log2\frac{Pr\ (x)}{1-Pr\ (x)},$$

where $Pr\ (x)$ is the probability of $(ys - ye)$ of x being greater than that of a random sequence (i.e., the significant level of silencing effect of x). The silencer alteration caused by a mutation was therefore measured as

$$\text{CNN-SAS-OR} = OR\_silencing(ref\ allele) - OR\_silencing(alt\ allele).$$

CNN-SAS-OR is highly correlated with CNN-SAS ($r = 0.6, p = 0$, Fig. S12B). The correlation of CNN-SAS-OR scores with raQTL scores on raQTLs is $r = -0.24$ ($p = 10^{-180}$, Fig. S12C), which is a close approximation to $r = -0.28$ of CNN-SAS scores ($p = 10^{-252}$, Fig. 3A).

TFBS prediction in TF ChIP-seq peaks

From the Encyclopedia of DNA elements project (https://www.encodeproject.org/), we downloaded TF ChIP-seq peaks reported for GM12878 lymphoblastoid cell line and H1 hESC cell lines. Here the peaks reported for GM12878 were used to approximate the binding events in T cells. The average length of the TF ChIP-seq peaks is around 300bp.

To decode and compare the binding compositions of DFREs for different functions, we predicted TFBSs within TF ChIP-seq peaks for each TF. Given a TF and the ChIP-seq peaks reported for this TF, we first derived the *de novo* motifs of the ChIP-seq peak sequences of the tested TF by using MEME CHIP (with the default setting) and HOMER findMotifsGenome.pl scripts (with the setting of -len 8,12,16 -size -100,100 -S 3). Among all *de novo* motifs, we then retained the one that was significantly enriched and had the highest abundance in the tested ChIP-seq peak sequences. The background sequences to derive the *de novo* motif were all sequences carrying the DNase-seq peaks or TF ChIP-seq peaks reported for the tested cell type. The mappings of the retained motif in TF ChIP-seq peaks were predicted as TFBSs of the tested TF. We applied this pipeline to each TF to predict its TFBSs.

TFBS prediction using a CNN model

The contribution of a nucleotide to silencer activity is evaluated as the average of the CNN output changes caused by all possible mutations on that nucleotide, i.e., $ds_{i,j}$.

$$ds_{i,j} = \frac{1}{3}\sum_{k=a,c,g,t}\left(ys_{i=k,j} - ys_{i=WT,j}\right) - \frac{1}{3}\sum_{k=a,c,g,t}\left(ye_{i=k,j} - ye_{i=WT,j}\right) \quad (1)$$

where i and j are the ith position in the silencer j. WT represents the wild-type genotype, while ys is the prediction of the CNN model on the probability of being a silencer. To smooth the curves of $ds_{i,j}$ and consider that the positions within binding sites have varied contribution to binding affinity, we used a 9 bp-wide window to screen the sequences with a sliding step of 1bp. In each window, the average of non-negative $ds_{i,j}$s (i.e., $wds_{i,j}$) was calculated. We retained the windows where the $wds_{i,j}$ values had significance $p < 0.1$ according to the empirical distributions of all windows in the silencers. The loose significance setting of $p < 0.1$ aimed to capture the marginal areas of binding sites. After merging the overlapping windows having a significant $wds_{i,j}$, we obtained the segments enriched with high $wds_{i,j}$s. Those regions, sensitive to the sequence mutations, were imputed as TFBSs. To identify the enhancer TFBSs, we used the average decrease of enhancer activity of all possible mutations, i.e., $-ds_{i,j}$ in eq. (1).

To evaluate the TFBS predictions, we assessed the coincidence of these TFBSs with TF ChIP-seq peaks reported in the corresponding cell types. A TFBS was regarded as coinciding with a TF ChIP-seq peak when 80% of its sequence overlapped a TF ChIP-seq peak. To address the limited resource for T cells, we used GM12878 TF ChIP-seq data to approximate the binding events in T cells. To mitigate the problem that long ChIP-seq peaks result in the inflated estimation about "coinciding", ChIP-seq peaks longer than 200bp were tailored into 200bp-long segments centering at the peak midpoints. As demonstrated in Fig. S17A, 49% of the predicted TFBSs in H1 hESCs coincide with TF ChIP-seq peaks profiled for H1 hESCs, which is 2 times that of randomly scattered TFBSs in the DFREs. In T cells, 43% of the predicted TFBSs coincide with TF ChIP-seq peaks for GM12878 (which was used as a proxy of T cells in this analysis), which is 1.4 times that of the randomly scattered TFBSs in the DFREs.

Also, we inspected the enrichment of published binding motifs in the predicted TFBSs, with the expectation that the predicted TFBSs are enriched for the binding motifs of the TFs essential for the tested biological context. The silencer TFBSs in the DFREs are enriched for the binding motifs of TCF4, SNAI1/2, and REST, among other repressors. On the other hand, the enhancer TFBSs in the DFREs show a high density of the binding motifs of hESC-specific TFs, such as POU5F1, NANOG, and SOX6. These results indirectly validated the CNN-based TFBS predictions. We downloaded sequence motifs associated with H3K27me3 (Ngo et al. 2019) and found that 81 used in our study match H3K27me3-associated motifs (using TOMTOM with default parameters). The motifs that are enriched in either DFREs, SLrs, or enhancers are included in Fig. S17B to demonstrate the sequence features of DFREs (and SLrs). Five of these motifs are H3K27me3-associated, which correspond to the TF motifs of SP110, MAX, USF1, SREBF2, and EHF.

Supplementary Figures

Supplementary Figure 1

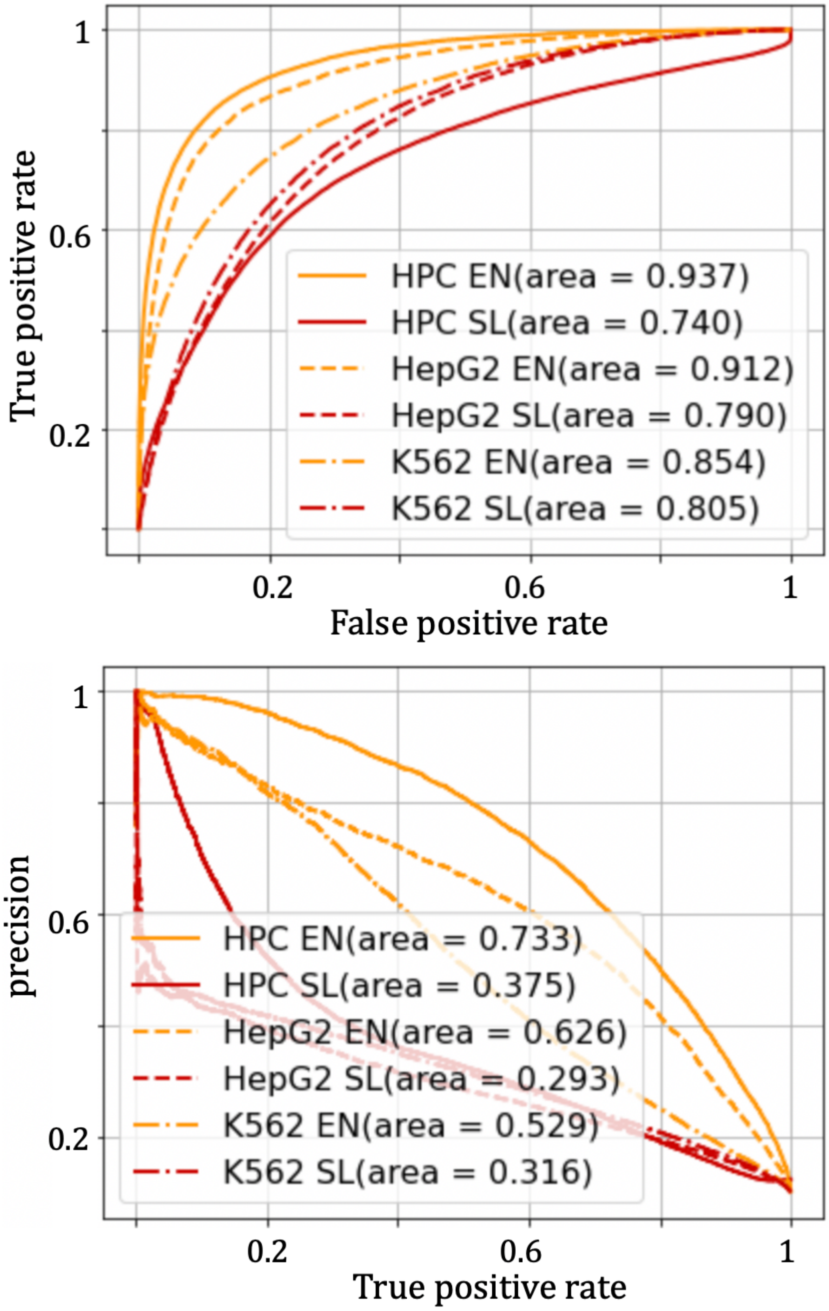


Figure S1. ROCs and PRCs of the HPC, HepG2, and K562 CNN models on test sequences.

Supplementary Figure 2

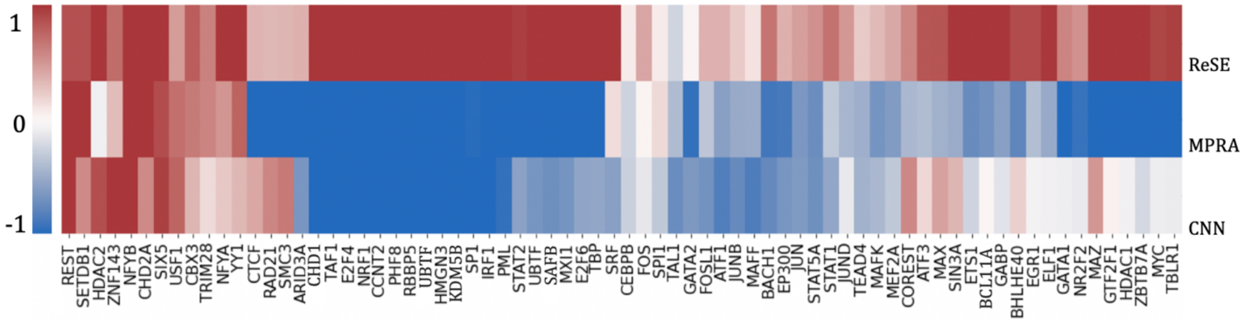


Figure S2. TF ChIP-seq peak enrichment fold (log_2) in different silencer sets reported for K562. CNN represents the silencers predicted by our CNN model. Background is all sequences harboring DNase-seq peaks or H3K27me3 ChIP-seq peaks detected in K562.

Supplementary Figure 3

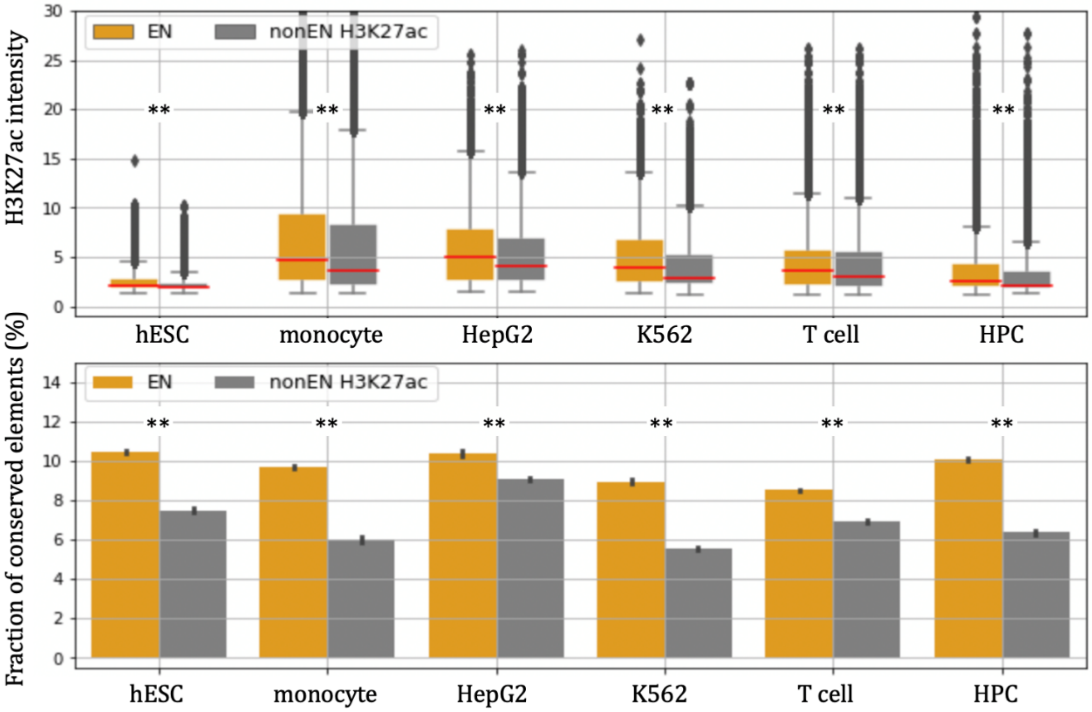


Figure S3. Comparative analysis of predicted enhancers (EN, orange) and the H3K27ac ChIP-seq peaks not predicted as enhancers (nonEN H3K27ac, grey) in terms of (a) H3K27ac signal intensity and (b) overlap with evolutionarily conserved elements across six examined cell types. In all these cell types, the predicted enhancers consistently carry stronger H3K27ac modifications (Student's *t*-test $p < 10^{-10}$) and a larger overlap with conserved elements than nonEN H3K27ac counterparts (Student's *t*-test $p < 10^{-10}$). $** - p < 10^{-10}$.

Supplementary Figure 4

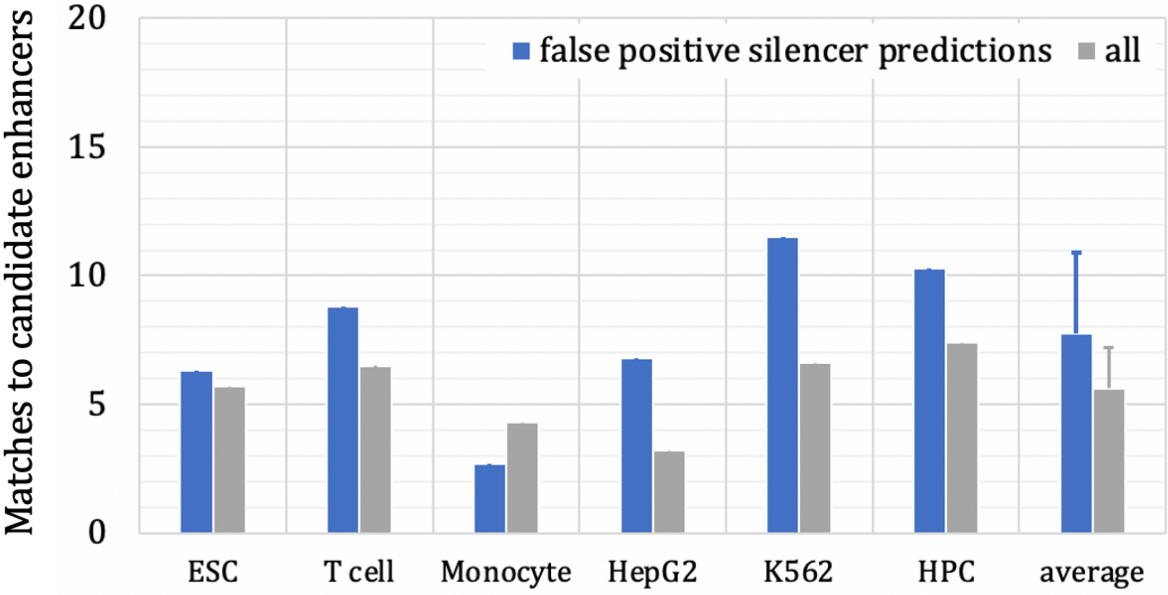


Figure S4. Fraction (%) of false positive silencer predictions (blue) matching candidate enhancers across six cell lines based on test sample data. False positive silencer predictions represent the sequences that carry no H3K27me3 signals but are predicted as silencers by the CNN model. The all (grey) represents all test sequences having no overlap with H3K27me3 ChIP-seq peaks.

Supplementary Figure 5

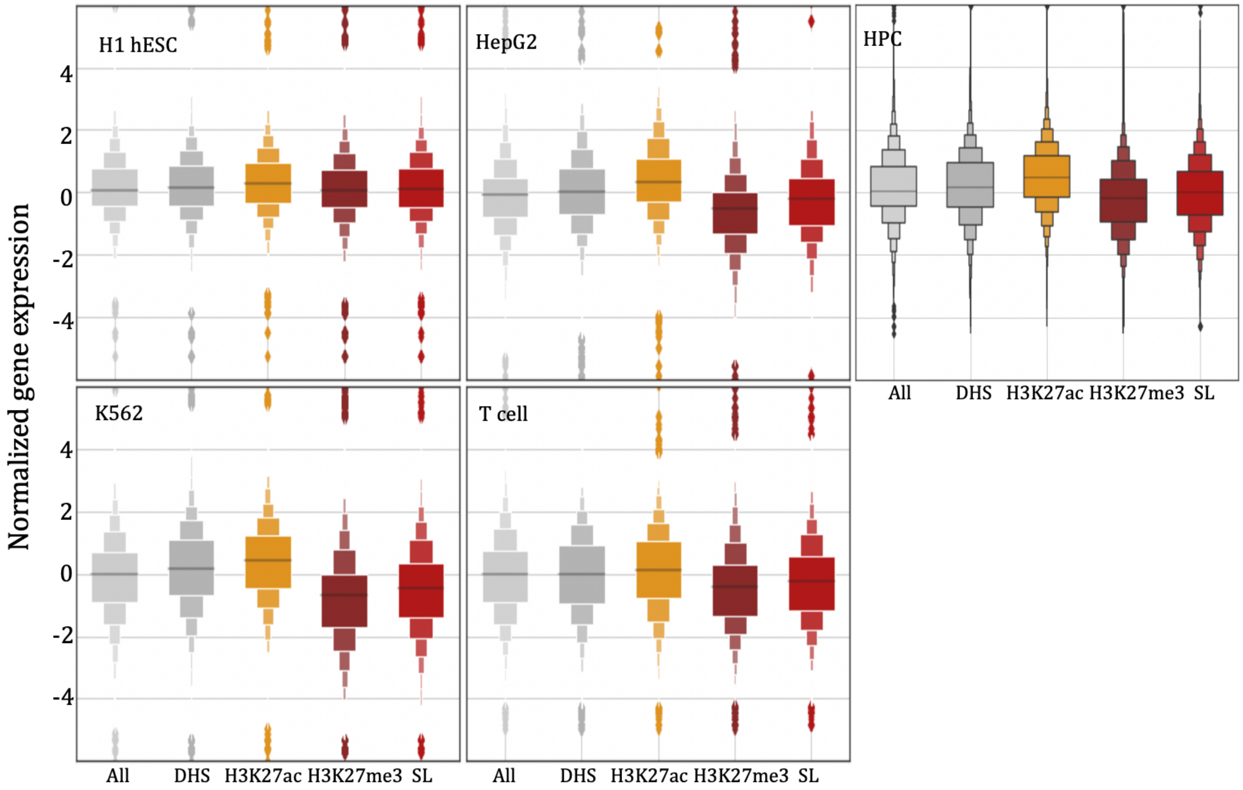


Figure S5. Distribution of normalized expressions of genes associated with different types of regulatory elements in different human cell types.

Supplementary Figure 6

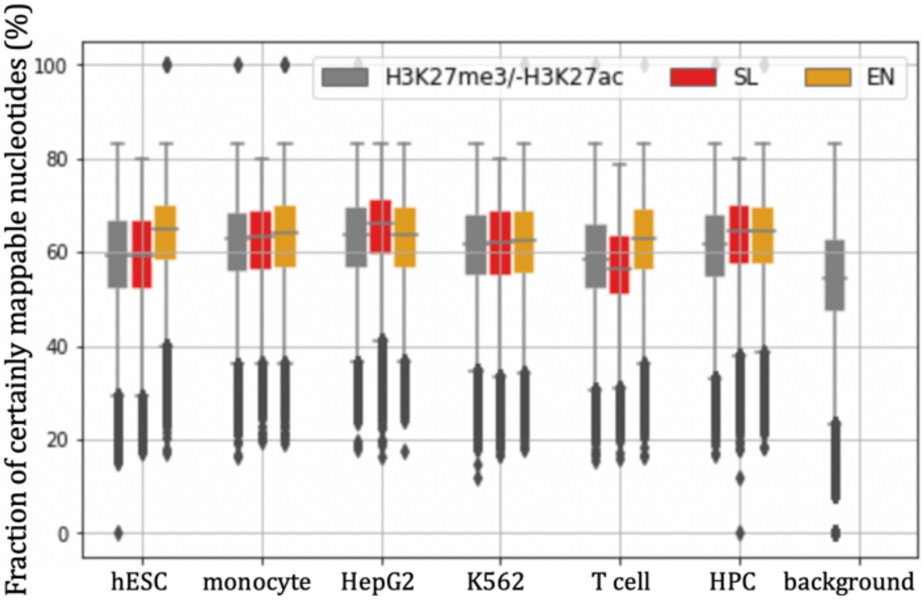


Figure S6. Fraction of certainly mappable nucleotides in genomic sequences. SL (red) and EN (orange) are predicted silencers and enhancers in the corresponding cell type, respectively.

Supplementary Figure 7

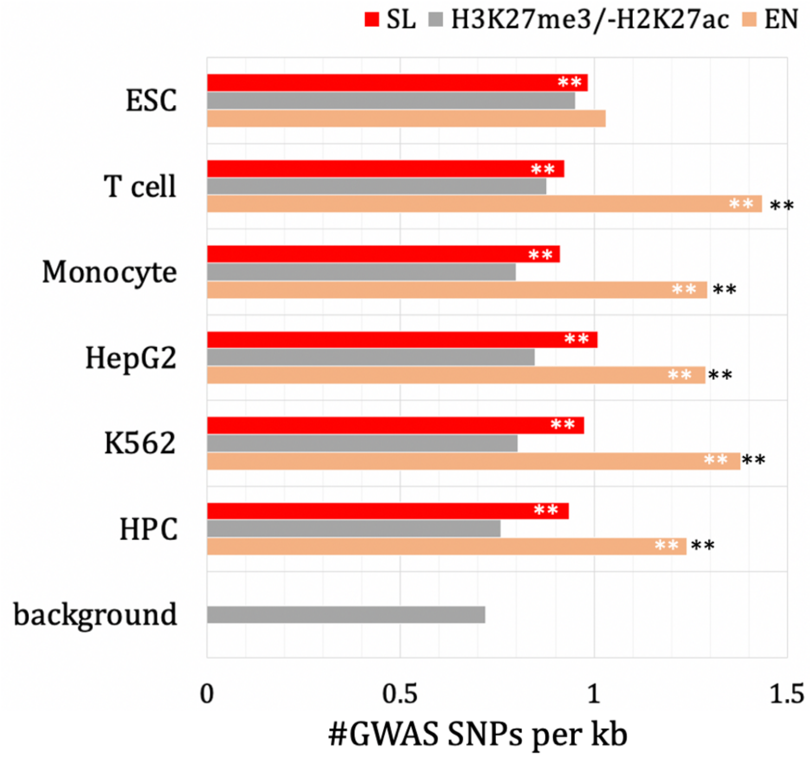


Figure S7. GWAS SNP densities of high-mappability sequences. SL (red) and EN (orange) denote high-mappability putative silencers and enhancers. H3K27me3/-H3K27ac (grey) denote high-mappability sequences with a H3K27me3 ChIP-seq peak but no H3K27ac ChIP-seq peaks. $** \ - p < 10^{-10}$. White asterisks represent significant enrichment of GWAS SNPs as compared to H3K27me3/-H3K27ac, and black asterisks are significant enrichment in high-mappability ENs as compared to the SL counterparts.

Supplementary Figure 8

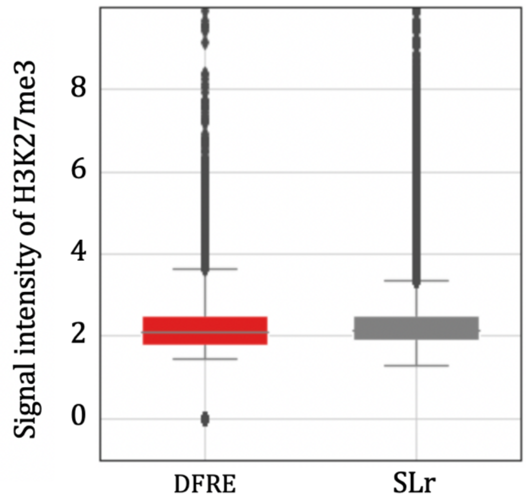


Figure S8. H3K27me3 modification intensity in DFREs (red) and SLrs (grey) in T cells.

Supplementary Figure 9

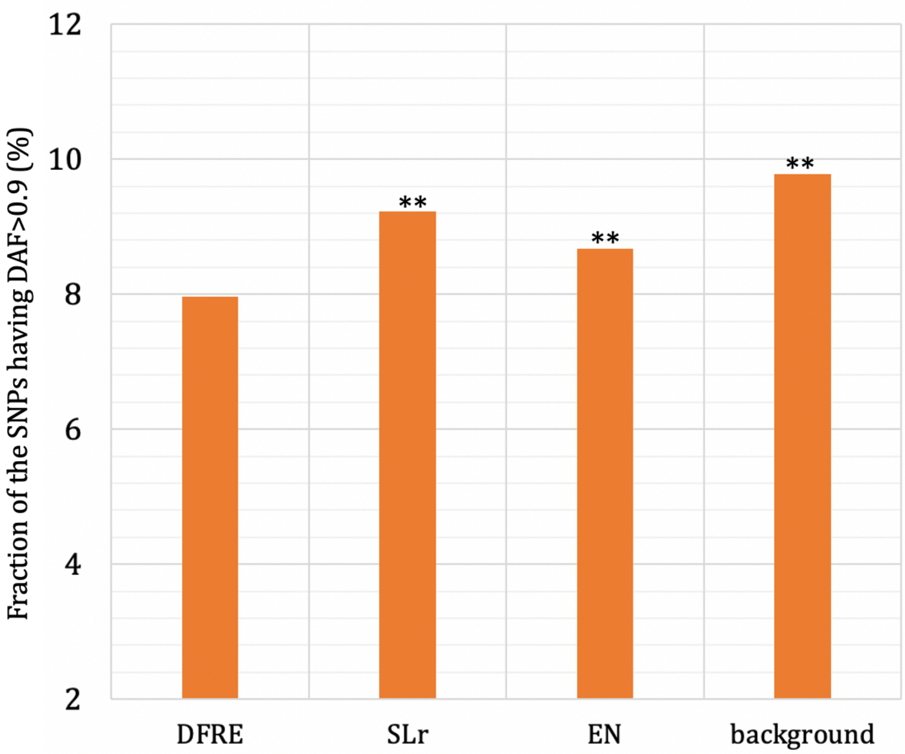


Figure S9. Fraction of the SNPs with derived allele frequency (DAF) >0.9. $** - p < 10^{-5}$, the enrichment significance levels as compared to the DFREs. EN represents the enhancers in T cells and background represents the sequences randomly selected from the human genome having the GC and repetitive element contents matching the silencers.

Supplementary Figure 10

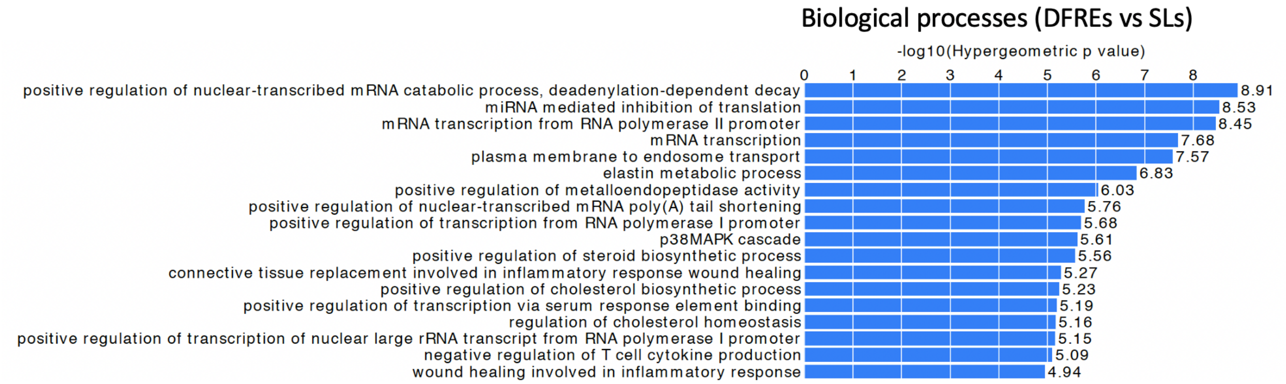


Figure S10. Biological processes that were significantly associated with the T cell DFREs. The results are from GREAT by using all T cell silencers (i.e., SLs) as background.

Supplementary Figure 11

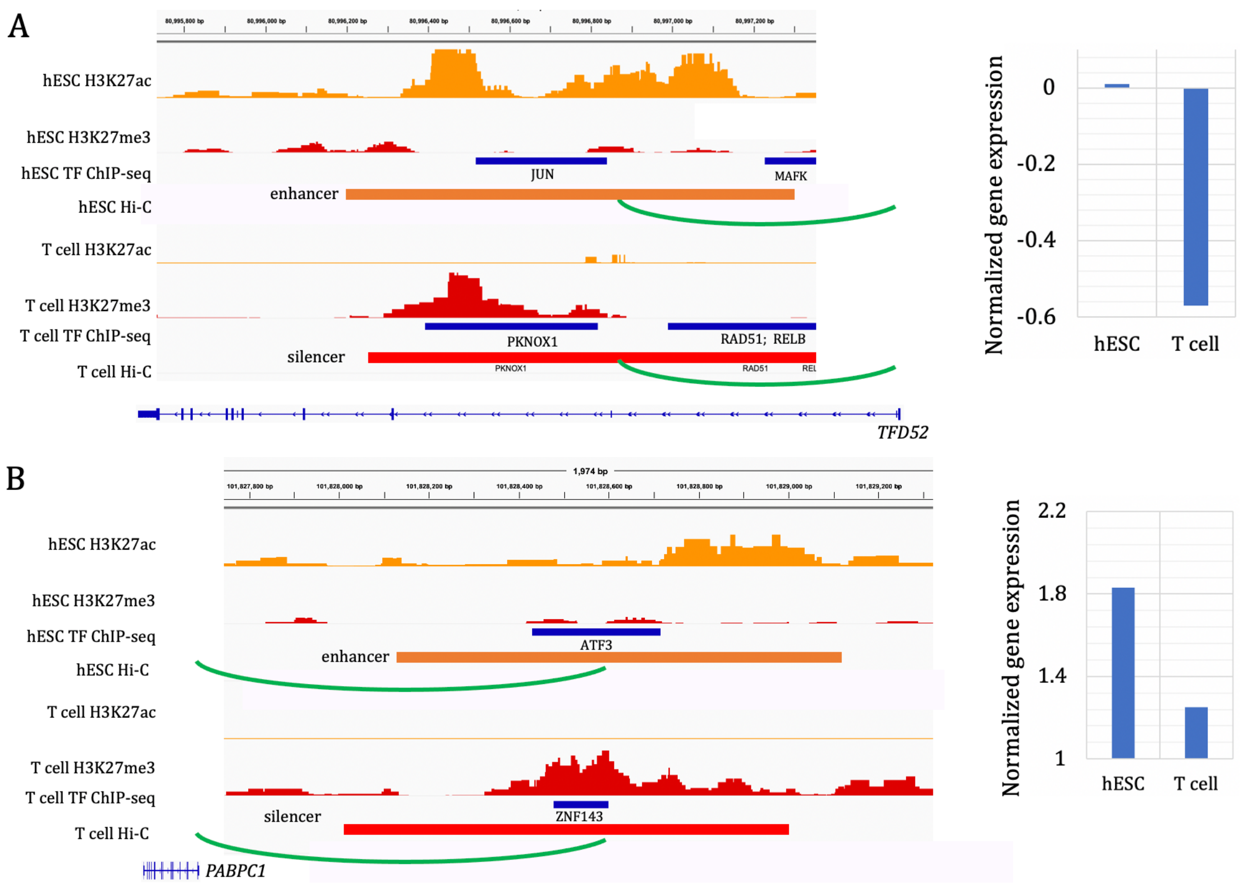


Figure S11. Examples of DFREs with Hi-C links to neighboring genes in both cell types. DFREs next to (A) *TFD52* and (B) *PABPC1*. The bar plots are the expression levels of the corresponding genes in the two cell types. The figures were generated using the Integrative Genomics Viewer (Robinson et al. 2011).

Supplementary Figure 12

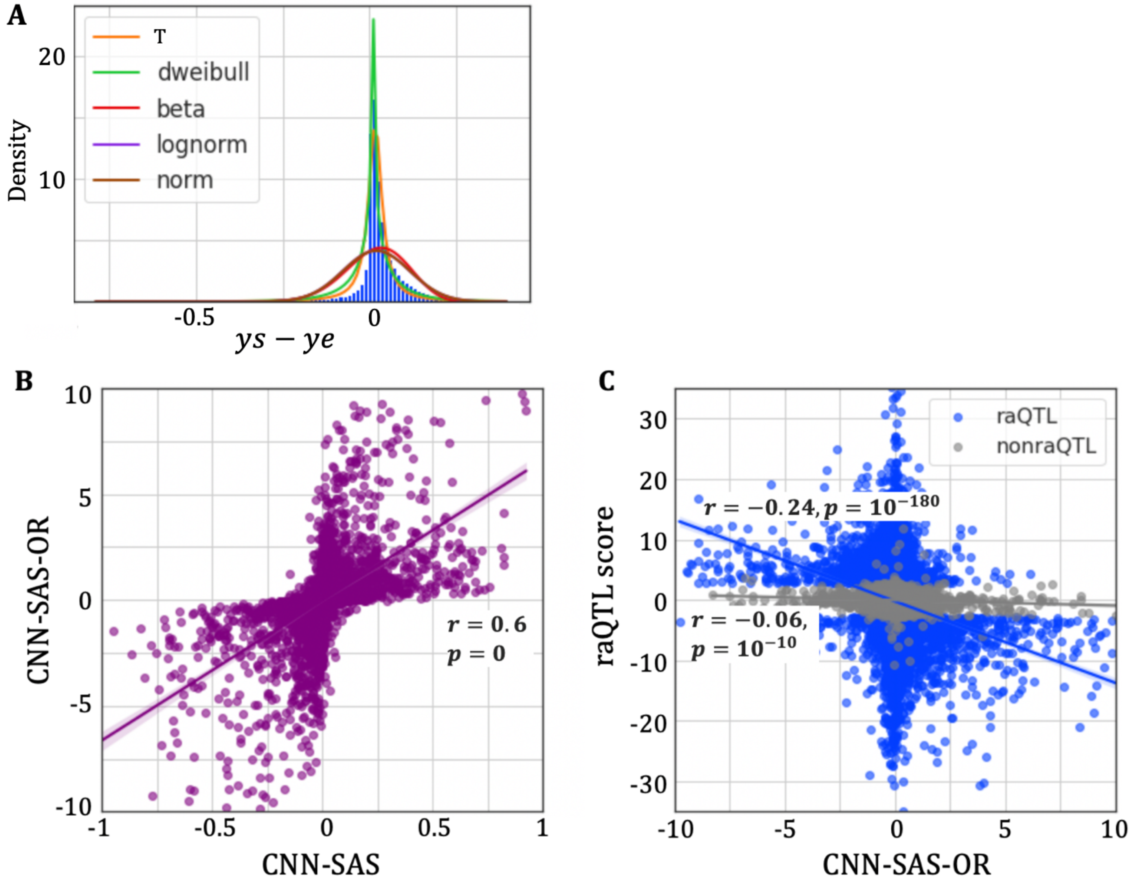


Figure S12. Comparison between CNN-SAS and CNN-SAS-OR scores. (A) Probability functions to fit the distribution of $ys - ye$. (B) High correlation between CNN-SAS and CNN-SAS-OR scores. (C) Correlation between CNN-SAS-OR and raQTL scores.

Supplementary Figure 13

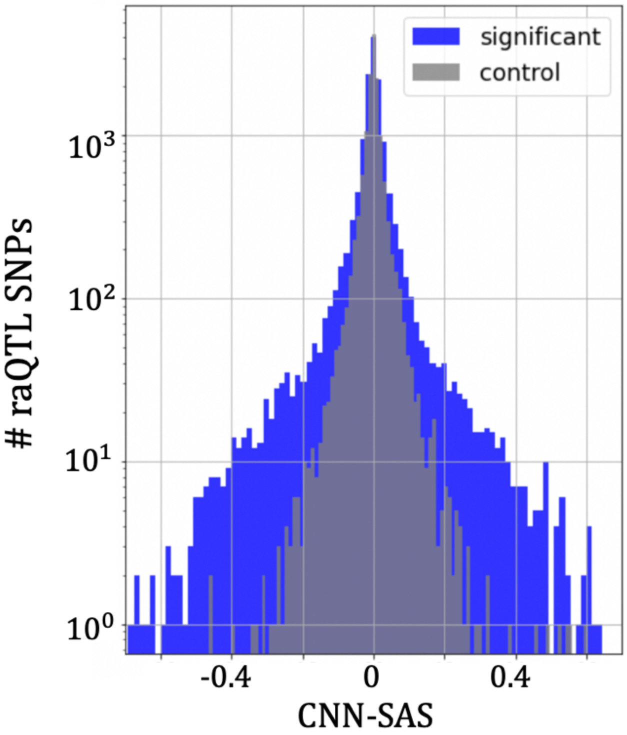


Figure S13. Distribution of CNN-SASs on the raQTL mutations (blue) in the HepG2 cell line. Control represents the non-raQTL mutations. Non-raQTL mutations were published along with raQTLs and have insignificant scores (grey).

Supplementary Figure 14

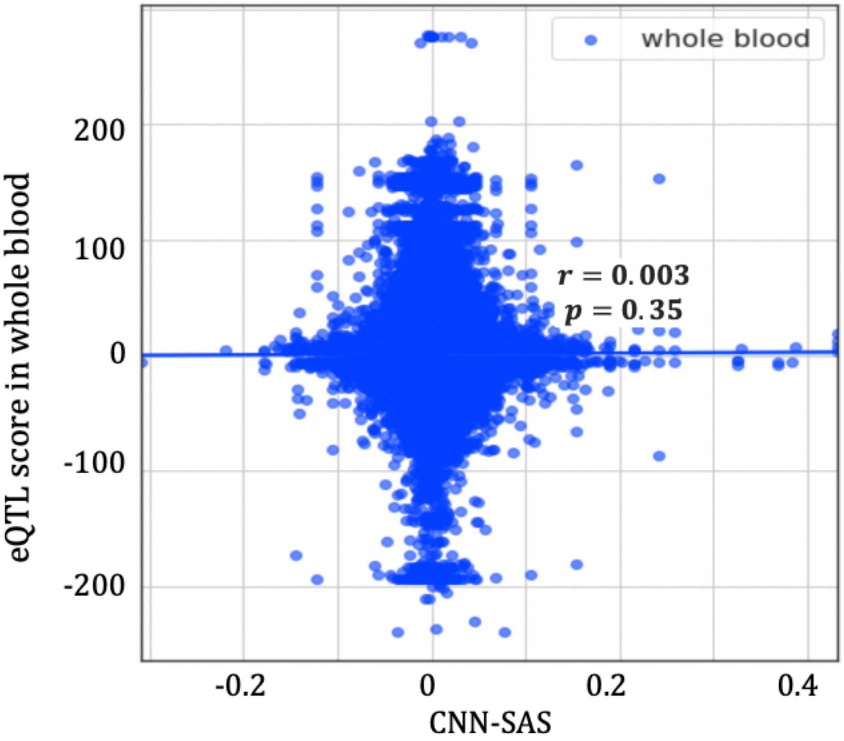


Figure S14. Correlation between CNN-SASs and eQTL scores detected in whole blood.

Supplementary Figure 15

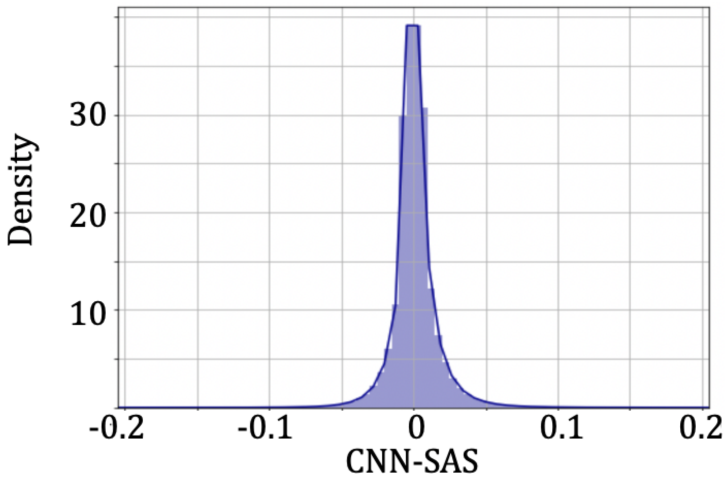


Figure S15. CNN-SAS scores of all possible single nucleotide silencer mutations in T cells. The solid line represents the estimate of a probability density function of the CNN-SASs, which was fitted by using seaborn kdeplot (Waskom 2021) with default parameters.

Supplementary Figure 16

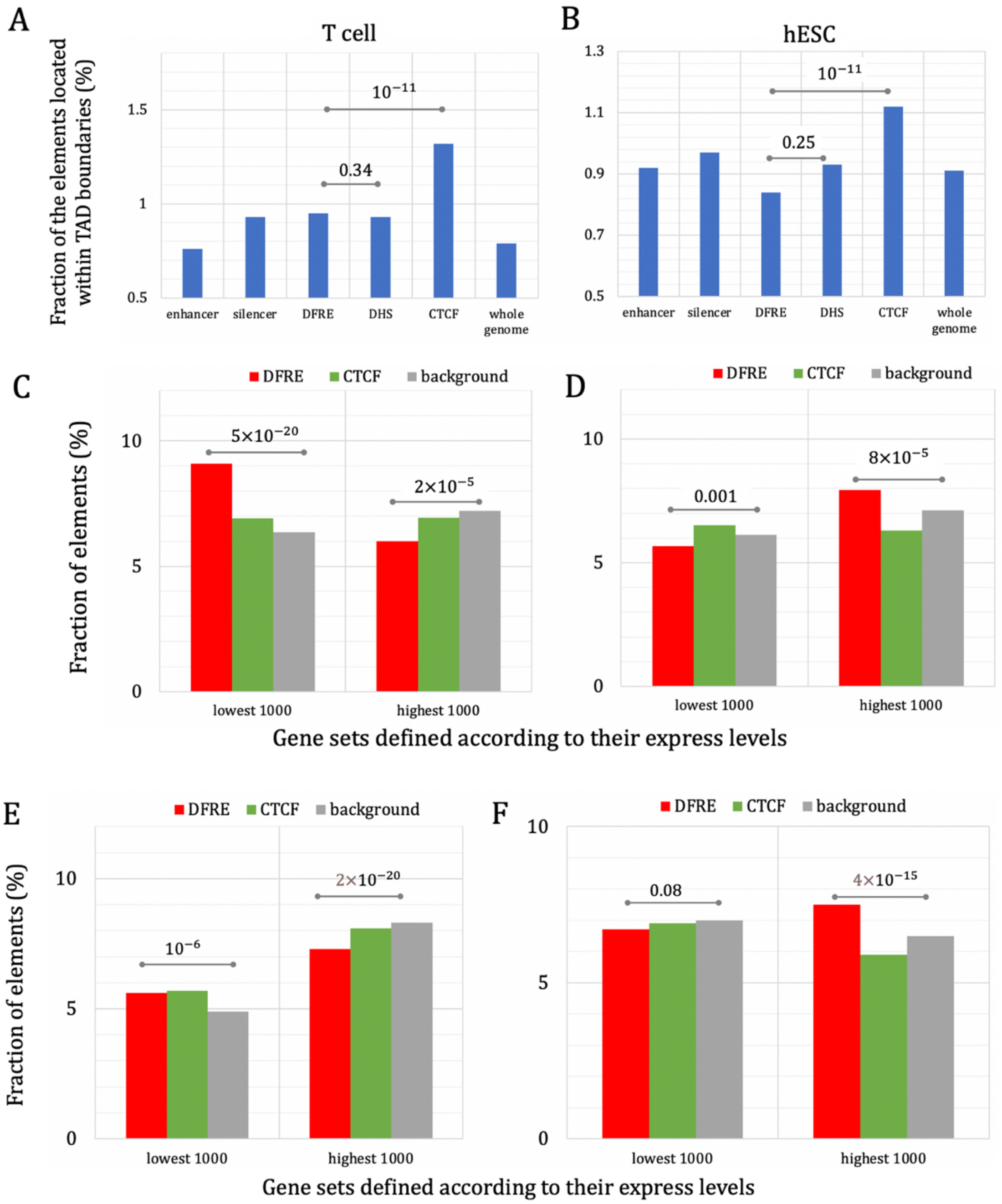


Figure S16. Comparisons between DFREs and CTCF-defined insulators. Fraction of elements located at the boundaries of TADs in (A) T cells and (B) hESCs. Fraction of elements located within the loci of the 1,000 lowest and highest expressed genes in (C) T cells and (D) hESCs. Backgrounds in (C) and (D) are the whole human genome. Fraction of elements contacting with the 1,000 lowest and highest expressed genes in (E) T cells and (F) hESCs. Backgrounds in (E) and (F) are all genomic regions carrying a DNase-seq peak or an H3K27me3 ChIP-seq peak.

Supplementary Figure 17

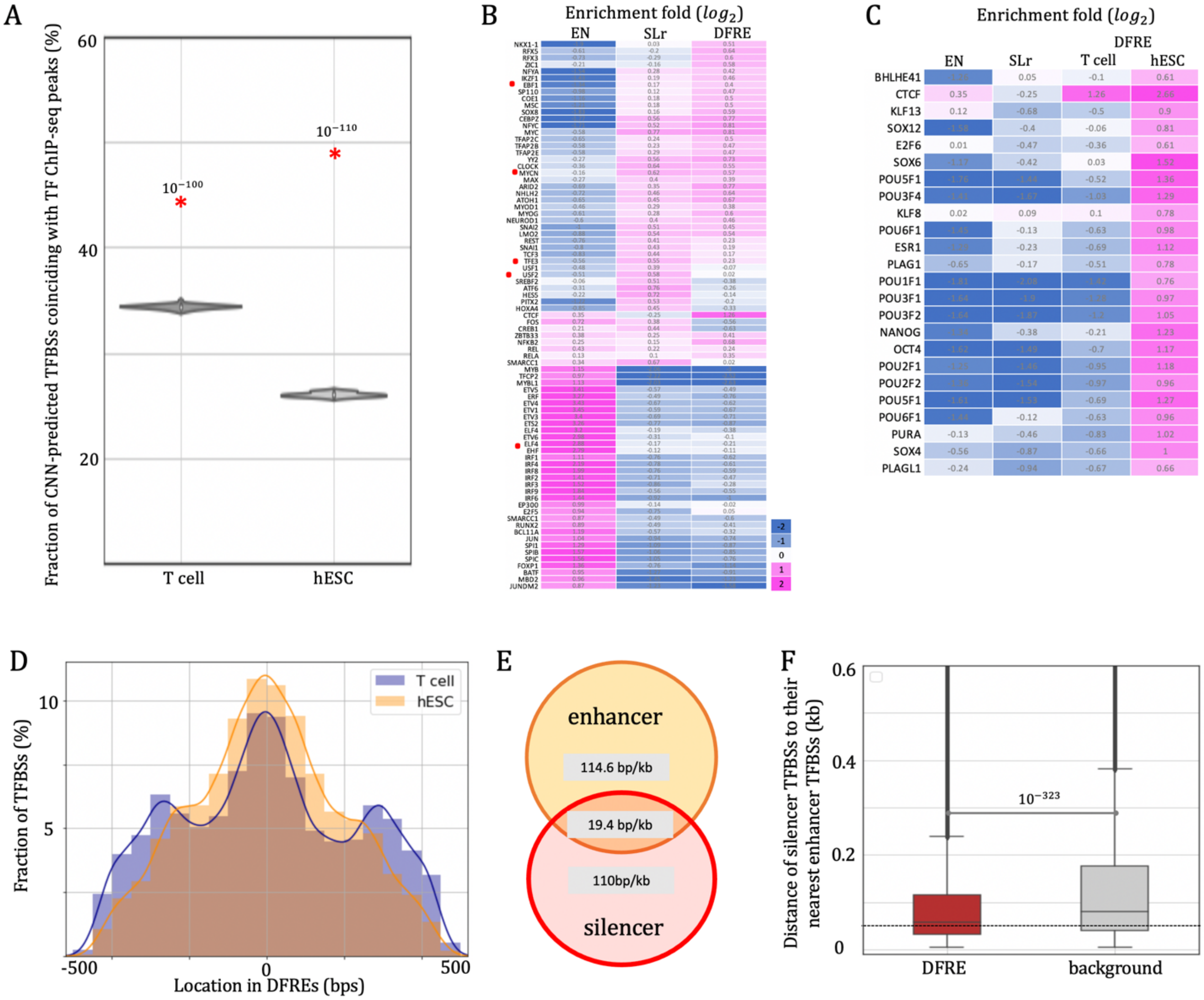
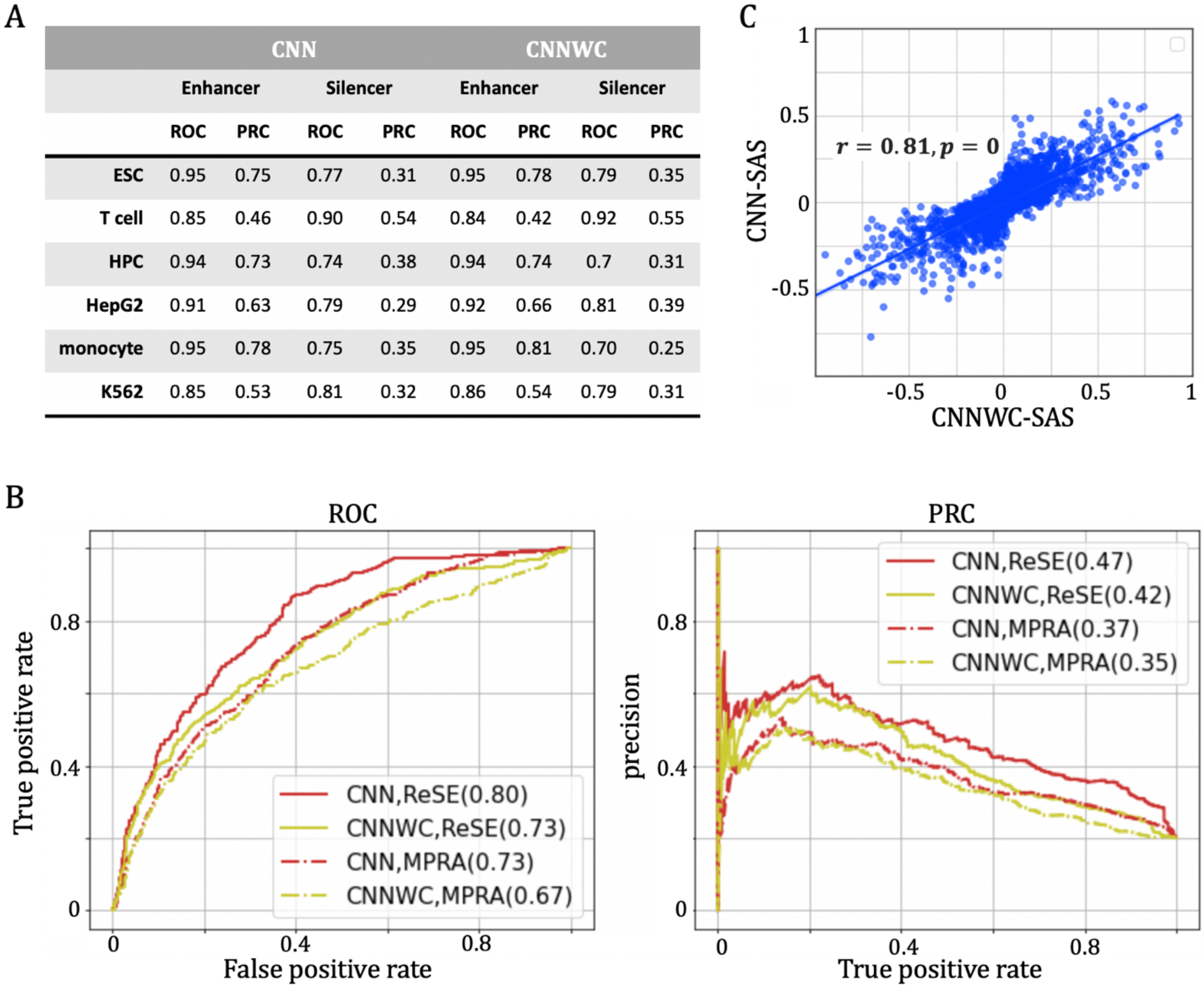


Figure S17. CNN-predicted TFBSs significantly coincides with TF ChIP-seq peaks in T cells and H1 hESCs. (A) Fraction of CNN-predicted TFBSs coinciding with TF ChIP-seq peaks. The red asterisks represent the CNN-predicted TFBSs (the numbers above are the significance p values). The gray violin plots are the background distributions estimated with the TFBSs randomly shuffled within DFRE sequences. TF ChIP-seq data in GM12878 were used to approximate the binding events in T cells, which potentially leads to the lower overlap between CNN-predicted TFBSs and TF ChIP-seq peaks in T cell than in H1 hESC. (B) TF motif enrichment in DFREs, ENs, and SLrs. Red dots mark the H3K27me3-associated motifs as presented in (Ngo et al. 2019). (C) TF motifs enriched in enhancer CNN-predicted TFBSs. Red dots mark the H3K27me3-associated motifs. (D) Distribution of silencer and enhancer TFBSs within DFREs. (E) Overlap between silencer and enhancer TFBSs within the DFREs. (F) Distance between silencer TFBSs to their nearest enhancer TFBSs. The background was generated through randomly shuffling CNN-predicted TFBSs.

Supplementary Figure 18



	CNN				CNNWC			
	Enhancer		Silencer		Enhancer		Silencer	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
ESC	0.95	0.75	0.77	0.31	0.95	0.78	0.79	0.35
T cell	0.85	0.46	0.90	0.54	0.84	0.42	0.92	0.55
HPC	0.94	0.73	0.74	0.38	0.94	0.74	0.7	0.31
HepG2	0.91	0.63	0.79	0.29	0.92	0.66	0.81	0.39
monocyte	0.95	0.78	0.75	0.35	0.95	0.81	0.70	0.25
K562	0.85	0.53	0.81	0.32	0.86	0.54	0.79	0.31

Figure S18. Comparison of the CNN models built using the original cost function (namely, CNN) with the models built using a weighted-class cost function (namely, CNNWC) in terms of classification performance (i.e., AUC ROC and AUC PRC) on (A) test samples and (B) experimentally validated silencers as well as in terms of (C) CNN-SASs on raQTLs.

Supplementary Table 1

Table S1. Predicted silencers and enhancers
A big table is given in a separate file supplemental_Table_S1.xlsx.zip

Supplementary Table 2

Table S2. GWAS traits associated with the T cell DFREs.

GWAS trait		#SNPs			enrichment fold		enrichment p value	
		DFRE	SLr	ENr	DFRE	SLr	DFRE	SLr
1	hepatitis B infection, Susceptibility to viral and mycobacterial infections	22	50	12	16.901	2.404	8.48E-20	6.08E-08
2	Granulomatosis with Polyangiitis	24	49	16	13.828	1.767	1.70E-19	0.000267
3	hepatitis B infection	25	237	23	10.021	5.946	4.96E-17	4.12E-100
4	airway imaging measurement	17	43	18	8.707	1.379	3.90E-11	0.039
5	chronic hepatitis B infection	27	245	29	8.5831	4.875	1.22E-16	3.34E-86
6	sensory perception of smell	15	107	21	6.585	2.94	2.11E-08	2.05E-21
7	chemerin measurement	21	24	30	6.453	0.461	5.11E-11	2.15E-05
8	acute graft vs. host disease	24	74	36	6.146	1.186	6.07E-12	0.145
9	Sjogren syndrome	25	119	38	6.065	1.807	2.96E-12	3.53E-09
10	oropharynx cancer	17	55	26	6.028	1.221	8.96E-09	0.136
11	hypothyroidism	17	150	32	4.898	2.705	1.66E-07	8.85E-26
12	response to vaccine	35	318	77	4.19	2.383	5.44E-12	7.08E-42
13	response to anticoagulant	23	93	56	3.786	0.958	1.24E-07	0.7224
14	HIV-1 infection, response to efavirenz, virologic response measurement	11	101	27	3.756	2.159	0.000239	5.05E-12
15	susceptibility to mumps measurement	16	74	40	3.688	1.068	1.31E-05	0.548
16	Tuberculosis	17	106	44	3.562	1.39	1.11E-05	0.0013
17	Ischemic stroke	16	109	46	3.207	1.367	6.74E-05	0.0018
18	Graves' disease	15	148	44	3.143	1.941	0.000138	2.97E-13
19	interleukin 18 measurement	5	124	15	3.073	4.77	0.025205	1.20E-43
20	lipoprotein A measurement	20	147	61	3.023	1.391	2.06E-05	0.00014
21	optic disc area measurement	17	140	52	3.014	1.554	8.53E-05	1.12E-06
22	susceptibility to shingles measurement	20	103	63	2.927	0.943	3.21E-05	0.5984
23	QRS amplitude, QRS complex	17	111	55	2.85	1.165	0.000163	0.112
24	pursuit maintenance gain measurement	23	255	75	2.827	1.962	1.47E-05	2.77E-22
25	vitamin D measurement	12	117	41	2.698	1.647	0.00218	5.39E-07
26	myocardial infarction	19	175	67	2.614	1.507	0.00021	3.07E-07

27	smoking status measurement, lung carcinoma	35	202	127	2.54	0.918	1.20E-06	0.2379
28	alcohol dependence	19	222	70	2.502	1.83	0.00035	2.44E-16
29	smoking behavior, unipolar depression	16	105	59	2.5	1.027	0.000991	0.7666
30	venous thromboembolism	20	228	76	2.426	1.731	0.000368	3.19E-14
31	Takayasu arteritis	16	213	61	2.418	2.015	0.001386	5.00E-20
32	fibrinogen measurement	23	236	93	2.28	1.464	0.000375	3.19E-08
33	HIV-1 infection	14	158	64	2.017	1.424	0.012977	2.28E-05
34	hip bone mineral density	18	118	85	1.952	0.801	0.007736	0.01497
35	metabolite measurement	22	186	105	1.931	1.022	0.00425	0.73872
36	monocyte percentage of leukocytes	52	552	250	1.918	1.274	1.88E-05	4.18E-08
37	Vitiligo	21	172	103	1.88	0.964	0.00657	0.6534
38	coronary heart disease	27	251	133	1.871	1.089	0.00329	0.17678
39	optic cup area measurement	13	156	65	1.844	1.385	0.0353	0.00011
40	optic disc size measurement	17	105	89	1.761	0.681	0.0336	3.27E-05
41	allergic rhinitis	14	167	74	1.744	1.302	0.0485	0.00106
42	attention deficit hyperactivity disorder	22	226	117	1.733	1.115	0.0157	0.10607
43	leukocyte count	48	515	256	1.729	1.161	0.00041	0.00089
44	alcohol use disorder measurement	32	175	173	1.705	0.584	0.00511	8.27E-15
45	lymphocyte percentage of leukocytes	55	404	299	1.695	0.78	0.00028	2.13E-07
46	neutrophil count, eosinophil count	29	318	158	1.692	1.161	0.00742	0.00855
47	primary biliary cirrhosis	45	266	246	1.686	0.624	0.00127	1.02E-16
48	basophil count, eosinophil count	43	435	237	1.673	1.059	0.0015	0.22668
49	alcohol consumption measurement	31	224	173	1.652	0.747	0.00768	5.66E-06
50	granulocyte count	29	324	162	1.65	1.154	0.01143	0.01116

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311-322.

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766-d773.

Johnson JM, Khoshgoftaar TM. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* **6**: 27.

Kim SG, Harwani M, Grama A, Chaterji S. 2016. EP-DNN: A Deep Neural Network-Based Global Enhancer Prediction Algorithm. *Scientific Reports* **6**: 38433.

Ngo V, Chen Z, Zhang K, Whitaker JW, Wang M, Wang W. 2019. Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. *Proceedings of the National Academy of Sciences* **116**: 3668.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**: 6.

Roadmap Epigenomics Consortium. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**: 24-26.

Singh AP, Mishra S, Jabin S. 2018. Sequence based prediction of enhancer regions from DNA random walk. *Scientific reports* **8**: 15912-15912.

Waskom ML. 2021. Seaborn: Statistical Data Visualization. *J Open Source Softw* **6**: 3021.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* **12**: 931-934.