SI Appendix

Supplementary Experimental Procedures

Preparation of positive and negative training sets

Positive training set

We selected the set of enhancer-promoter (EP) pairs based on a recently published ChIA-PET data set. Using an anti-RNA polymerase II antibody (1), Li *et al.* identified chromatin interactions involving promoters in human K562 and MCF-7 cells. Not all of these interactions are between promoters and enhancers. Thus, we first identified enhancers in these two cell types using the CSI-ANN algorithm (2) and three histone modifications that together uniquely mark active enhancers (H3K4me1, H3K4me3, and H3K27ac). Our current knowledge about chromatin marks for enhancers is incomplete. There are additional chromatin marks such as H4K16ac and H2A.Z. However, it is generally believed the combination of above three marks is by far the minimal combination that give accurate prediction of active enhancers (3). Next, we used the following stringent criteria to select enhancers that overlap with reported ChIA-PET interactions: 1) *cis* interactions with ≥5 PET counts (more stringent than the 3 counts used by the authors); 2) one interacting site contains p300 site (enhancer marker) but not promoter, and the other interacting site contains promoter but not p300 site; 3) promoters need to be expressed based on matching RNA-Seq data (i.e. RPKM value > 0). The inventors of ChIA-PET developed a statistical analysis framework to account for the random formation of any inter-ligation PETs between two anchors such that both inter-ligation PET frequency and ChIP enrichment of the anchors are taken into account (4). Based on this model, predicted interactions with three or more inter-ligation PETs between anchors have a false discovery rate < 0.05. Thus, PET >= 3 was used as the cutoff for calling interactions. Here, we chose a more stringent cutoff of >=5

PETs to ensure the better quality of our training data.

Using the above set of stringent criteria, we extracted 1124 and 1110 enhancer and promoter pairs for K562 and MCF7 cells, respectively. The selected enhancers have higher levels of histone marks and the selected target promoters have higher expression levels in the corresponding cell types (Fig. S2), further supporting the quality of these interactions.

Negative training set

A naïve approach to selecting negative training pairs is to randomly select a promoter for a given enhancer. However, the contact frequency between two non-interacting genomic loci in a chromatin fiber does not follow a uniform distribution. Instead, it is a function of the site separation distance in the following form (5):

$$\boldsymbol{f}(\mathbf{s}) = \boldsymbol{k} \times \boldsymbol{s}^{-3/2} \times \boldsymbol{e}^{-1400/s^2}$$

where *s* denotes the sites separation distance, and the proportionality constant *k* reflects the efficiency of the cross-linking reaction. In our analysis, to generate a set of non-interacting EP pairs, for each enhancer, we first randomly selected a site based on the contact frequency distribution described above. Then we selected the closest promoter to that site as the candidate target. We also ensured that the selected promoter was not detected by ChIA-PET (i.e. < 3 PET). Otherwise, we would use the next closest promoter to the site until it met our criteria. As a result, we selected a non-interacting promoter for each of all 2234 enhancers in the positive training set.

Histone modification ChIP-Seq and RNA-Seq data

The ENCODE consortium has generated genome-wide histone modification maps and gene expression profiles for multiple human cell lines. In this analysis, we collected histone modification and RNA-Seq data from ENCODE for the following eleven cell lines: GM12878, H1 ES, HepG2, HMEC, HSMM, HUVEC, IMR90, K562, MCF-7, NHEK, and NHLF. In addition, we collected previous published histone modification and RNA-Seq data for CD4$^+$ T cell (6, 7).

Annotation of known transcripts and promoters

We defined promoter region as upstream 2 kbp to downstream 0.5 kbp of annotated transcription start site as defined by GENCODE (8). Enhancer region is defined as a 2 kbp window predicted by CSI-ANN (2).

Compendium of transcription factor motifs

Transcription factor motifs were obtained from the Jaspar (9), TRANSFAC (10), and Uniprobe (11) databases. Redundant motifs were removed by manual inspection.

Construction of Receiver Operating Characteristic (ROC) Curve.

ROC curves were used to evaluate the performance of the methods based on training set (Fig. 1E) and external data sets (Fig. 2B-D). Given a set of predictions and a gold-standard set (either training set or external data sets), the following quantities were defined: True Positives (TP), predicted EP pairs that were supported by interactions in the gold-standard set; A predicted EP pair is considered to be true positive if the center of the enhancer in the predicted pair falls within one of the genomic regions of the gold-standard pair and the TSS in the predicted pair falls within the other genomic region of the gold-standard pair; False Positives (FP), predicted EP pairs that were not supported by interactions in the gold-standard set; False Negatives (FN), EP pairs that were not predicted but are found in the gold-standard set; and True Negatives (TN), EP pairs that were not predicted and are not found in the gold-standard set. True positive rate (TPR) was defined as TP/(TP+FN) and false positive rate (FPR) was defined as FP/(FP+TN). The curve is generated by computing TPR and FPR values on prediction sets derived by varying classifier decision threshold.

Genome-wide prediction of EP pairs

We first predicted enhancers genome-wide across in 12 cell types using CSI-ANN and the following 3 histone modifications: H3K4me1, H3K4me3, and H3K27ac. These histone marks are reported to be associated with active enhancers and commonly used to predict enhancers (12-14). Next, for each enhancer, we extracted all promoters within the 2 Mbp window centered at the enhancer. For each candidate EP pair within the window, we computed the feature scores of EPC, TPC, COEV, and DIS. Feature scores were combined in the RF model and a linkage score was computed for each candidate EP pair. We used False Discovery Rate (FDR) to set cutoff for making predictions. We computed the FDR using the training set. Specifically, FDR was defined as the fraction of training set pairs above a given linkage score threshold that are from the negative training set. We examined published ChIA-PET, 5C and high-resolution Hi-C datasets and found that the average number of target promoters per enhancer reported in the literature ranges from 2 to 6. Based on this observation, we set the final FDR cutoff to 1%, which yields on average 2.92 targets per enhancer across the 12 cell types.

Overlap of predicted enhancers with other genomic marks

Three types of genomic features were used to evaluate the predicted enhancers: DHS sites, p300 sites, and evolutionary conservation. For p300 and DHS sites, the data are in the form of ChIP-Seq peaks. They were downloaded from ENCODE. An enhancer prediction is considered to be supported by a given genomic feature if the center of a feature peak is located within the 2 kbp enhancer region. For sequence conservation, an enhancer prediction is considered to be conserved if 10% of its sequence (200 bp) has a phastCons conservation score > 0.5. By using 0.5 as the cutoff, approximately 5% of the human genome is conserved across the set of 35 placental mammalian genomes.

Calculation of transcript expression specificity rank

To calculate the expression specificity of a transcript, we compiled a compendium of RNA-Seq expression profiles from eleven human cell types. Following (15, 16), we calculated a specificity score for each transcript using an entropy-based measure that quantifies the skewness of expression level toward a given cell type. Briefly, given $\boldsymbol{N}$ cell types, we define the relative expression of transcript $\boldsymbol{g}$ in cell type $\boldsymbol{t}$ as:

$$\boldsymbol{p_{t|g} = w_{g,t}/\sum_{t=1}^{N} w_{g,t}}$$

where $\boldsymbol{w_{g,t}}$ is the expression level of transcript $\boldsymbol{g}$ in cell type $\boldsymbol{t}$. The entropy of a transcript's expression distribution is

$$\boldsymbol{H_g = \sum_{t=1}^{N} -p_{t|g} log_2(p_{t|g})}$$

To measure the expression specificity of a transcript, we first computed

$$\boldsymbol{Q_{g|t} = H_g - log_2(p_{t|g})}$$

A small $\boldsymbol{Q_{g|t}}$ indicates high expression specificty of transcript *g* in cell type *t*. Using this measure, we then ranked all transcripts in given cell type and computed a normalized expression specificity (referred as expression specificity rank in the manuscript) by dividing the rank with the total number of transcripts in the given cell type.

Comments on random forest model complexity and its relationship to training set size

Training of a Random Forest classifier involves building a set of decision trees, each of which is trained on a different random subset of the training dataset and a random subset of the available features is used to choose how best to partition the dataset at each node. Commonly used rule for deciding on the number of random features used for building each tree is the square root of the number of features available. Thus we used 2 randomly selected features for training each component tree of the forest. But for training the overall RF model, all four features were used.

The randomness introduced by the random forest model builder in the dataset selection and in the feature selection delivers considerable robustness to noise, outliers, and over-fitting, when compared to a single tree classifier. Because many trees are built and there are two levels of randomness and each tree is effectively an independent model, the model builder tends not to overfit to the training dataset. It is also proven that Random Forest classifier performance does not degrade as the number of trees increases (17).

The specific RF model in this study consists of a thousand trees. However, we also have a lot of training data, i.e. 4 types of features and each feature has more than 4000 training data points (~1000 positive EP pairs each from K562 and MCF-7 cells and matched number of negative training pairs). Thus, we do not think there is an overfitting issue. To further rule out this possibility, we also tested models with a range of RF parameter settings, including the number of trees, number of features in each tree, and maximum depth of the trees. For each RF parameter setting, we also test three 5 cross validation schemes: 5-fold, 10-fold, Train-K and Test-M, Train-M and Test-K, and randomized positive set. Here “Train-K and Test-M” means training with K562 positive and negative EP pairs and testing MCF-7 positive and negative pairs. “Train-M and Test-K” means the other way around. By using training and testing from different cell types, we rule out the possibility that the classifier’s good performance is not due to its “memory” of training data. For “randomized positive set” approach, the correct EP links in the positive set was destroyed/randomized, i.e. pairing the same set of promoters in the original positive set with a set of random enhancers with matched spatial distribution and histone marks. We then used the trained RF classifiers to predict the real positive set of EP pairs. In this scheme, because the classifier is deliberately fed with scrambled training data, we expect that the performance will be low. Please see Fig. S4 and S5, and Table S2.

External ChIA-PET, Hi-C, and eQTL-gene pair datasets for evaluate genome-wide EP predictions

We downloaded reported eQTL-gene associations from the University of Chicago eQTL browser (http://eqtl.uchicago.edu/cgibin/gbrowse/eqtl/). There are 4970 non-redundant pairs in liver cells from 2 studies (18, 19), and 87570 non-redundant pairs in lymphoblastoid cells from 8 studies (20-26).

We downloaded reported ChIA-PET interactions in K562 and MCF-7 cells from (1) and $CD4^+$ T cell from (27).

We downloaded reported Hi-C interactions in IMR90 cell from (28).

Performance comparison to methods by Ernst et al., Thurman et al. and PreSTIGE

Since predicted EP pairs were not provided by the authors, following the description in Ernst *et al.* (12), we implemented a logistic regression classifier using distance and EPC as features. For a given enhancer, when making predictions, we only considered promoters within 125 kbp of the enhancer as candidate targets as done in Ernst *et al*.

Thurman *et al*. (29) predicted EP pairs based on their DHS signal correlation. We downloaded the set of predicted EP pairs provided by the authors. The authors used 0.7 as the cutoff for predictions in their study. For performance comparison, we examined predictions using different DHS correlation thresholds, 0.7, 0.8, and 0.9.

PreSTIGE predicted EP pairs by pairing cell type-specific H3K4me1 signals with genes that are specifically expressed in each cell type across a panel of diverse cell types. The authors made two sets of predictions, high- and low-confidence sets. We downloaded the set of predicted

EP pairs provided by the authors. For EP pairs in IMR90 cell, we used the web interface to the PreSTIGE method to make prediction, using H3K4me1 ChIP-Seq and RNA-Seq data from ENCODE.

We benchmarked the performance of various methods on predicting genome-wide EP pairs using the three external datasets described above. We used each of the three external datasets as the gold-standard to construct ROC curves and compute F1 scores.

Assumptions of statistical tests

All statistical tests were performed using large sample sizes. Sample sizes were reported in figure legends. Other assumptions of specific tests such as normal distribution for t-test were tested to be satisfied before conducting the real tests. Therefore, the test results are robust with regard to underlying assumptions of the statistical tests.

Identification of CNC and CAC sites overlapping the predicted EP pairs

We downloaded genome-wide peaks of CTCF and cohesin subunits (SMC3, RAD21, and STAG1) identified by ENCODE for five cell types (GM12878, H1 ESC, HepG2, K562, and MCF-7). We defined SMC3/STAG1 peaks or RAD21 peaks that do not overlap with a CTCF peak as CNC sites, otherwise CAC sites. The overlapping criterion is peak center-to-center distance less than half of the length of the longer peak. Afterwards, we overlapped the CNC and CAC sites with the predicted EP elements to determine the role of CNC and CAC sites in mediating EP interaction.

Cell culture

GM12878 (cat. no. GM12878) and K562 (cat. no. GM05372) cells were purchased from Coriell Institute for Medical Research. Cell lines were tested for mycoplasma contamination using ABI MycoSEQ mycoplasma detection assay (Applied Biosystems). GM12878 cells were grown in

RPMI 1640 medium (Gibco) supplemented with 15% Fetal Bovine Serum (FBS) (Gibco), penicillin (Invitrogen), streptomycin (Invitrogen). K562 cells were grown in RPMI 1640 medium (Gibco) supplemented with 10% Fetal Bovine Serum (FBS) (Gibco), penicillin (Invitrogen), streptomycin (Invitrogen). Cultures were seeded at a concentration of between 200,000 and 500,000 viable cells per mL. Medium was changed every 2-3 days depending on cell density. Cells were harvested at log phase for 3C-qPCR experiments.

Supplementary Figures

Fig. S1. Flow chart for the selection of training set of EP pairs and training of the Random Forest classifier.

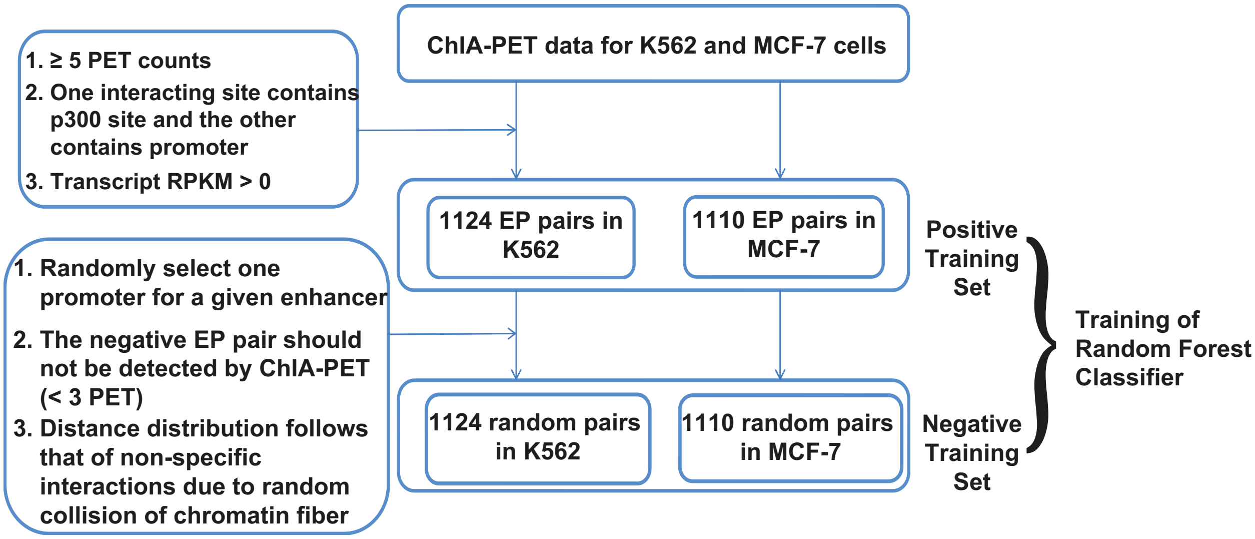

ChIA-PET data for K562 and MCF-7 cells
1. ≥ 5 PET counts
2. One interacting site contains p300 site and the other contains promoter
3. Transcript RPKM > 0
1124 EP pairs in K562
1110 EP pairs in MCF-7
Positive Training Set
1. Randomly select one promoter for a given enhancer
2. The negative EP pair should not be detected by ChIA-PET (< 3 PET)
3. Distance distribution follows that of non-specific interactions due to random collision of chromatin fiber
1124 random pairs in K562
1110 random pairs in MCF-7
Negative Training Set
Training of Random Forest Classifier

Fig. S2. (A) Boxplot of transcript levels of K562 promoters of the training set EP pairs. In comparison, expression levels of the same set of promoters in ten other cell types are shown. **(B)** Boxplots of transcript levels of selected MCF-7 promoters of training set EP pairs and the same set of promoters in ten other cell types are shown. **(C)** Boxplot of CSI-ANN prediction scores of selected K562 enhancers of the training set EP pairs. In comparison, CSI-ANN scores of the same set of enhancer sequences in ten other cell types are shown. **(D)** CSI-ANN prediction score of selected MCF-7 enhancers of the training set EP pairs. In comparison, CSI-ANN scores of the same set of enhancer sequences in ten other cell types are shown.

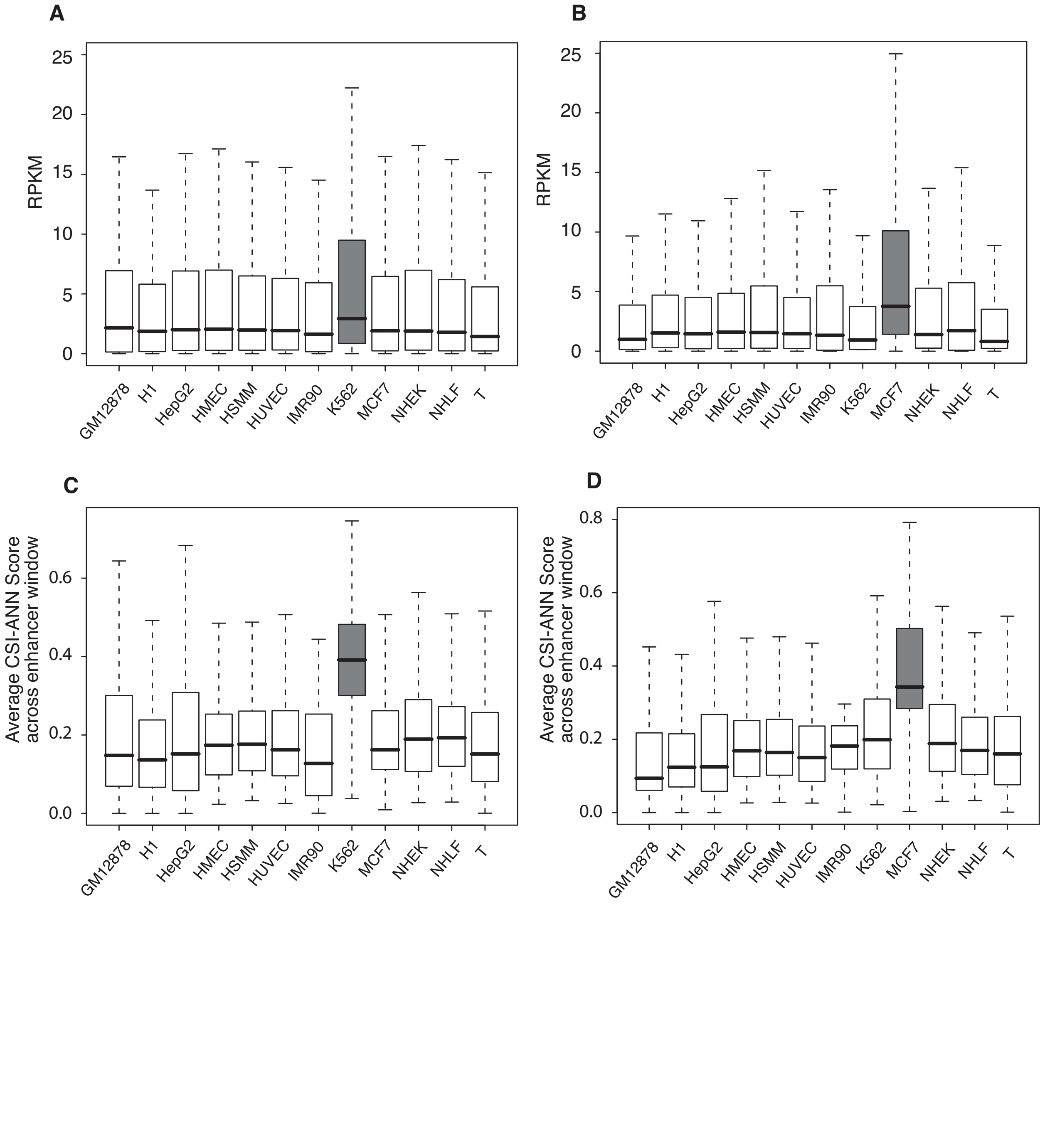


Fig. S3. Distance distribution of positive training set EP pairs in (A) K562 cell and **(B)** MCF-7 cell. Red lines are fitted probability density functions of geometric distributions. Goodness-of-fit p-values are shown in the figure.

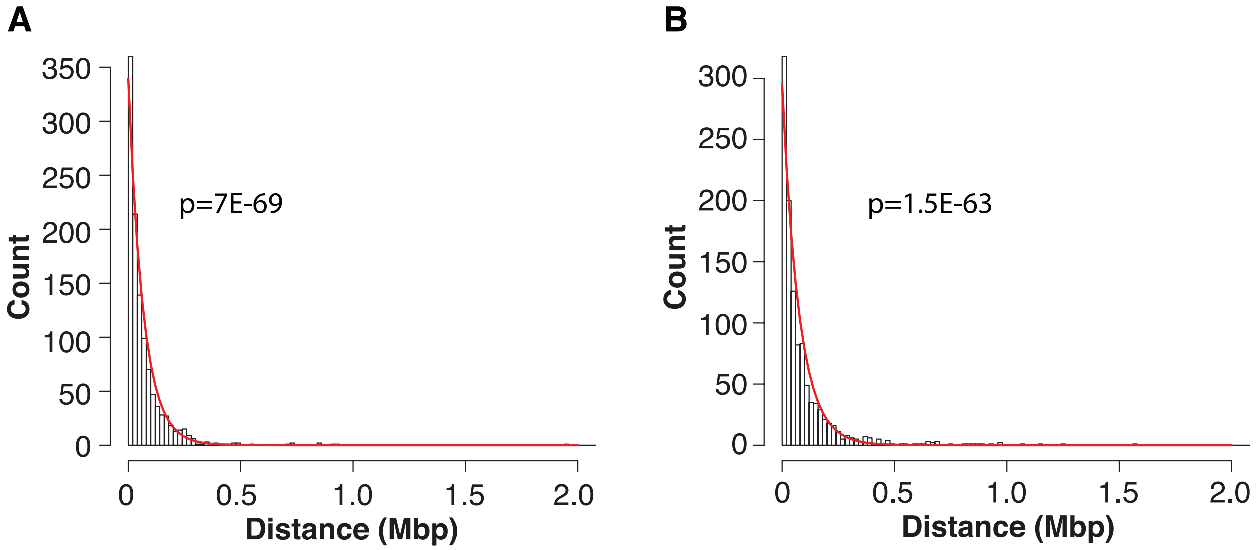


Fig. S4. Receiver operating characteristic curves for IM-PET trained using scrambled training data. Correct EP links in the positive set was destroyed/randomized, i.e. pairing the same set of promoters in the original positive set with a set of random enhancers with matched spatial distribution and histone marks. We then used the trained RF classifiers to predict the real positive set of EP pairs. We used three types of training-testing schemes, 5-fold cross validation, train using K562 data and test using MCF-7 data, train using MCF-7 data and test using K562 data. By The AUC for all three experiments are around 0.5, suggesting that the true positive set pairs cannot be correctly predicted using randomized training data.

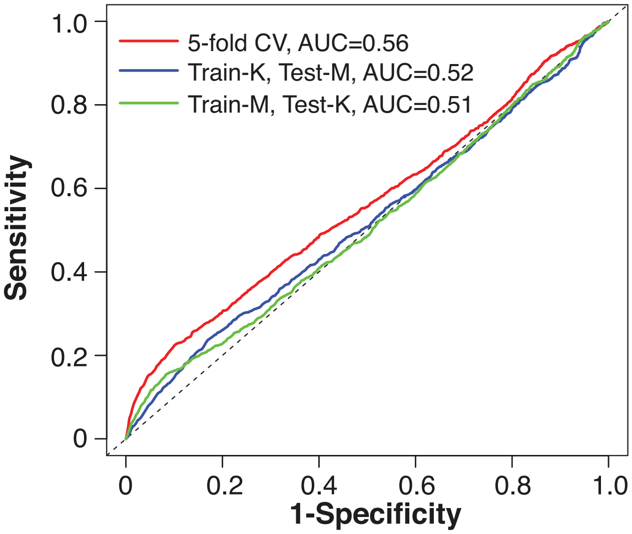


Fig. S5. Receiver operating characteristic curves for IM-PET using different classifiers and human data. RF4, random forest classifier using four features; RF2, random forest classifier using two features EPC and DIS; LR4, logistic regression classifier using four features; SVM4, support vector machine classifier using four features.

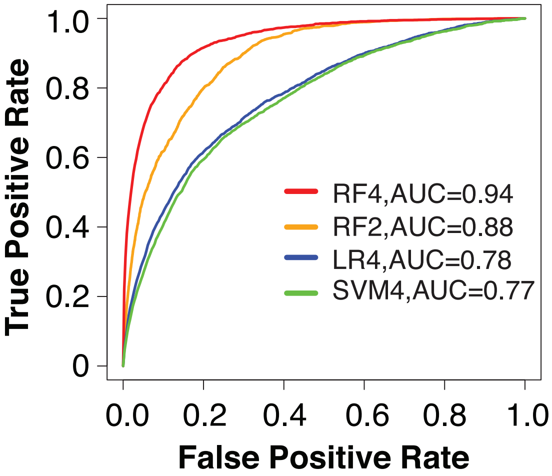


Fig. S6. Discriminative features and performance evaluation of the IM-PET algorithm applied to fruit fly datasets. (A) Enhancer and target promoter activity profile correlation (EPC); **(B)** TF and target promoter expression correlation (TPC); **(C)** Co-evolution of enhancer and target promoter (COEV); **(D)** Distance constraint between enhancer and target promoter (DIS); **(E)** Receiver operating characteristic curve. P-values are based on one-sided Student's t-test. N= 831 for all tests. RF4, random forest classifier using four features; LR2, logistic regression classifier using two features as used in Ernst et al.; Nearest-promoter, the approach of assigning the promoter(s) nearest to an enhancer as its target(s).

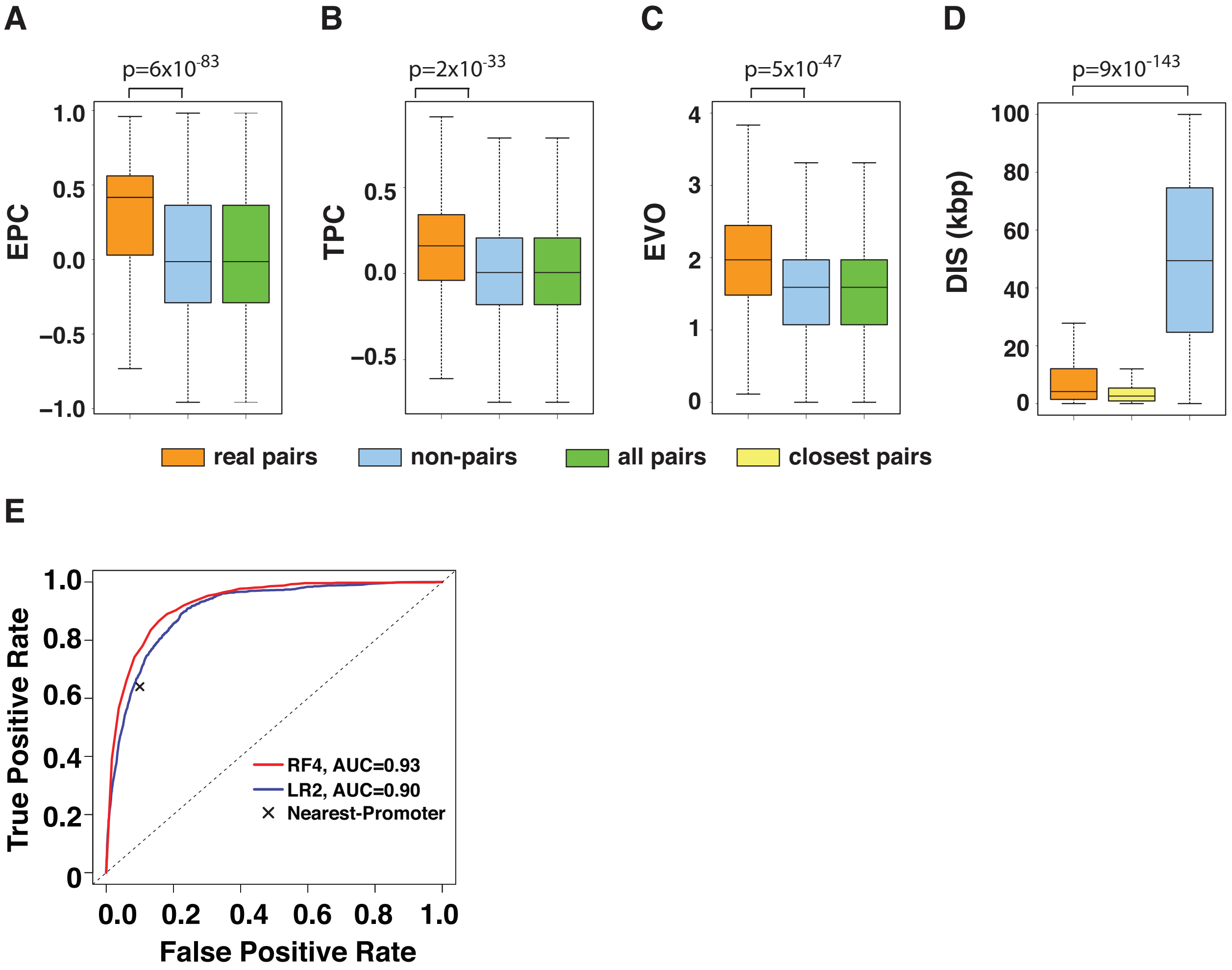


Fig. S7. 3C-qPCR validation at the *DDX39B* locus in GM12878 cell. (A) Genome browser view of predicted EP pair. The following tracks are shown from top to bottom: 3C-qPCR primer positions for negative controls (blue) and test (red) interactions; Refseq gene and transcript IDs (black) of the locus being tested; p300 ChIP-Seq peak; DHS ChIP-Seq peak; H3K4me1 ChIP-Seq peak; H3K27me3 ChIP-Seq peak. **(B)** Calibrations to identify the linear range for qPCR on BAC clone DNA control; **(C)** Calibrations to identify the linear range for qPCR on 3C DNA template; **(D)** The 3C result confirms the interaction at *DDX39B*. The EP pair is predicted in both K562 cell and GM12878 cell.

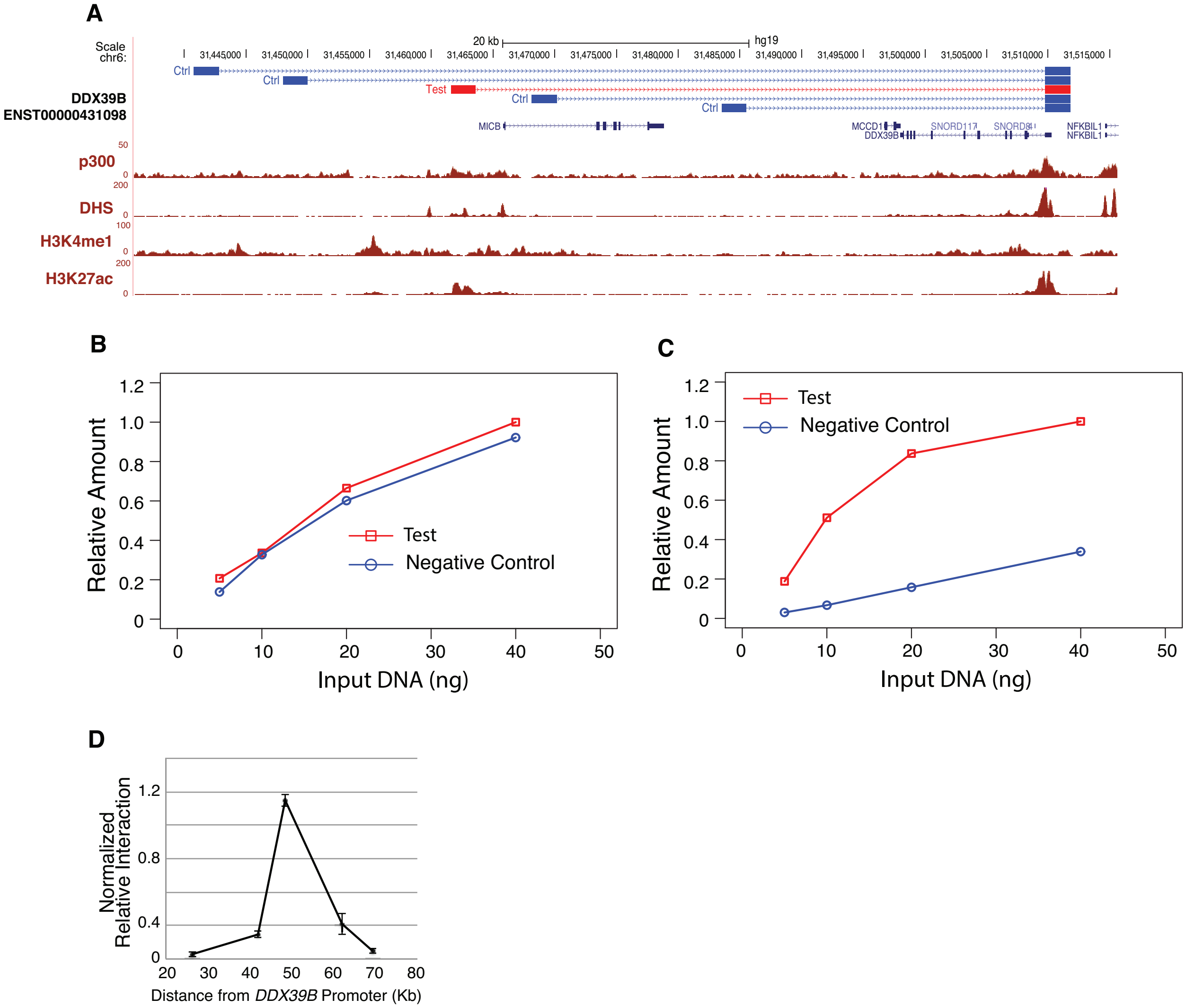


Fig. S8. 3C-qPCR validation at the *CD53* locus in GM12878 cell. The EP pair is predicted in GM12878 cell but not K562 cell.

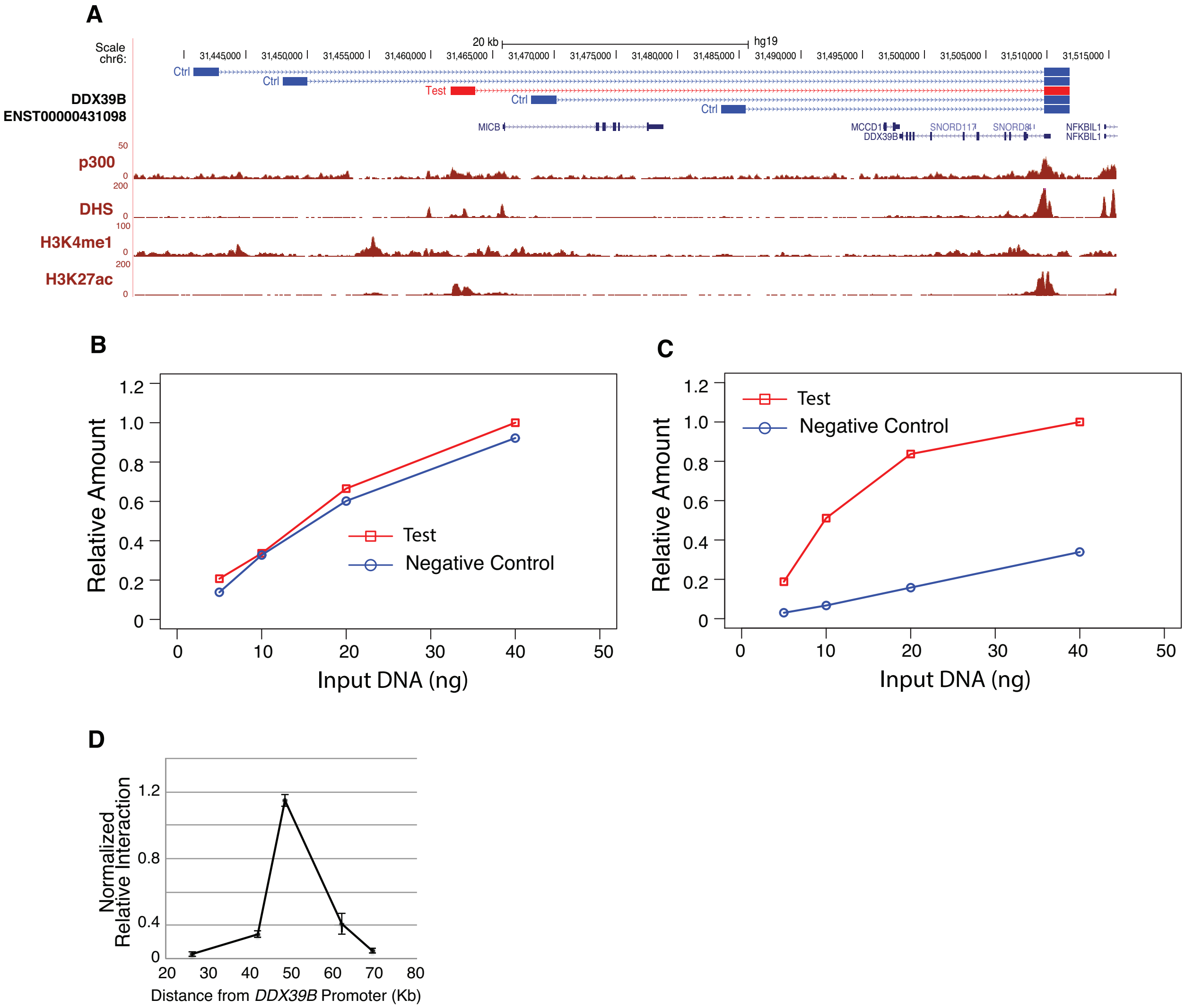


Fig. S9. 3C-qPCR validation at the *POU2AF1* locus in GM12878 cell. The EP pair is predicted in GM12878 cell but not K562 cell.

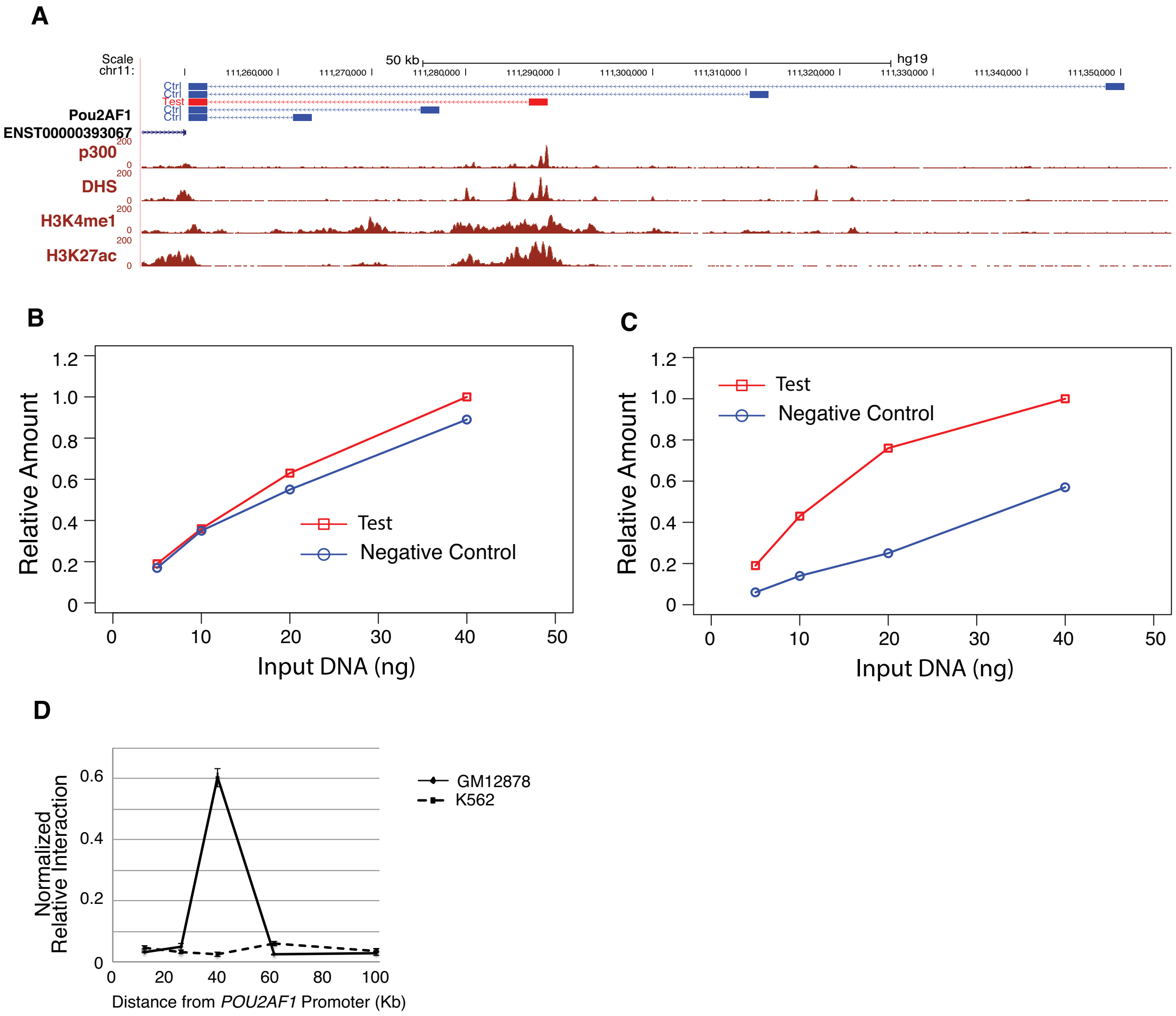


Fig. S10. 3C-qPCR validation at the *DDX39B* locus in K562 cell. The EP pair is predicted in both K562 cell and GM12878 cell.

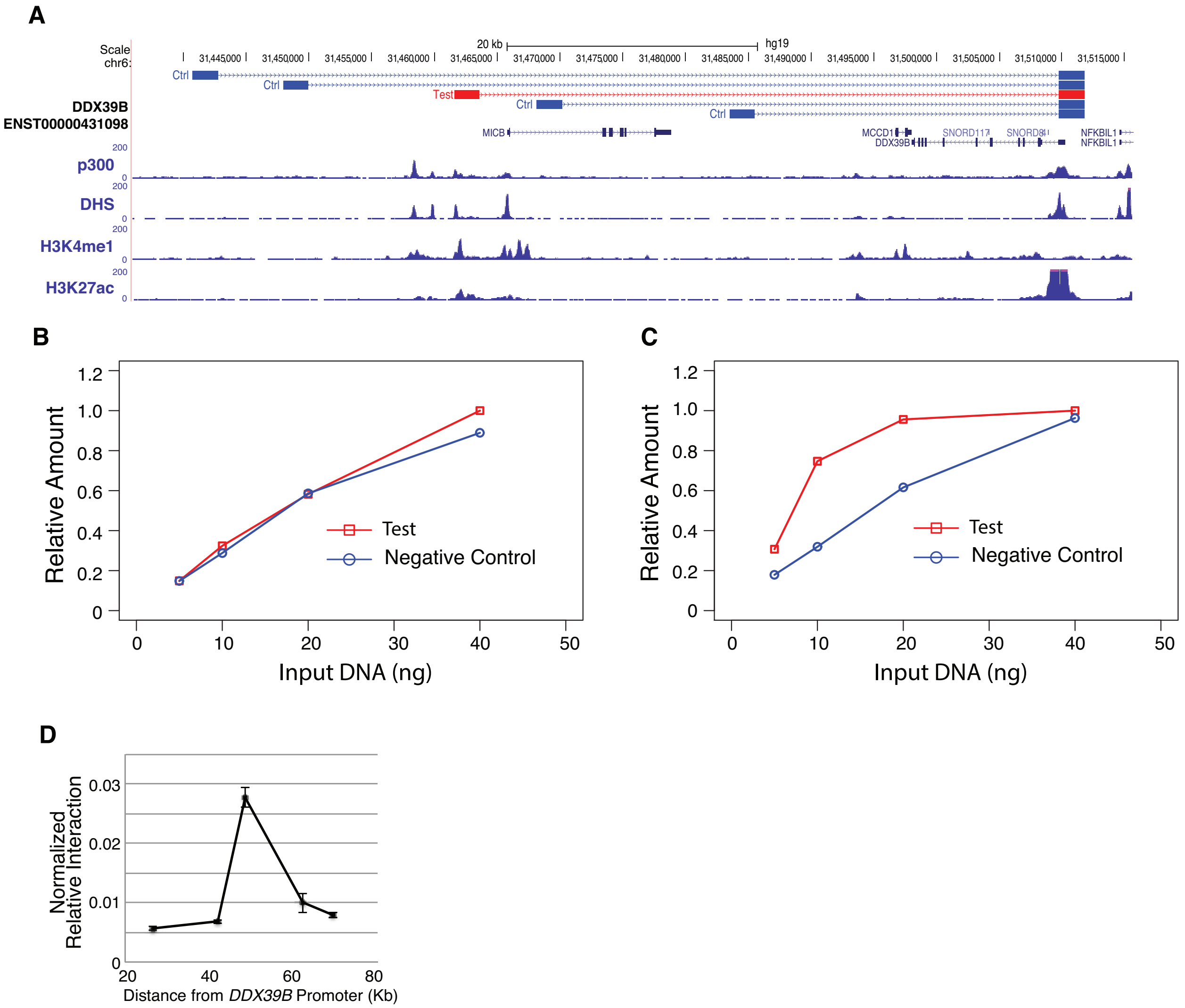


Figure S11. 3C-qPCR validation at the *GTSF1* locus in K562 cell. The EP pair is predicted in K562 cell but not GM12878 cell.

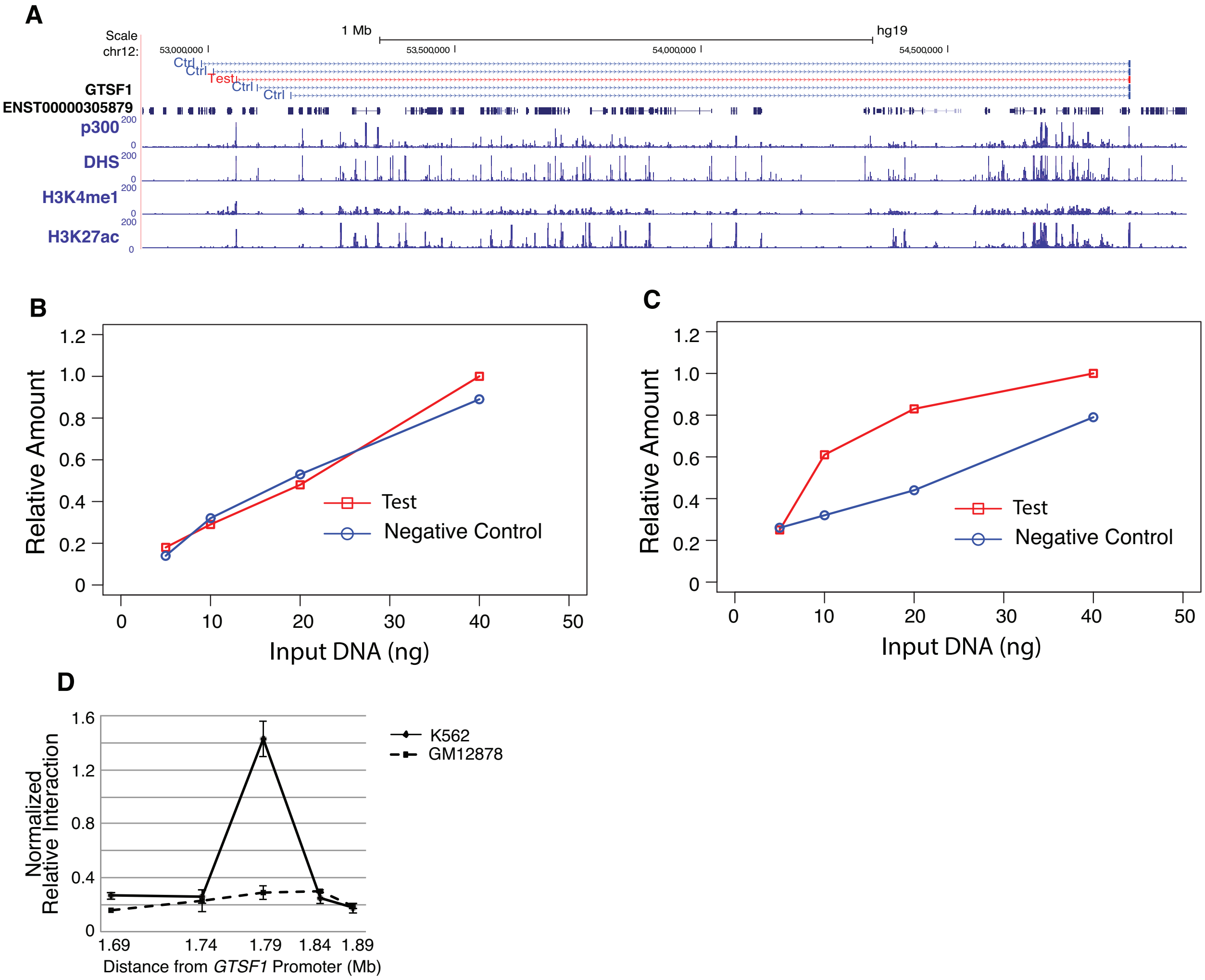


Fig. S12. 3C-qPCR validation at the *PEAR1* locus in K562 cell. The EP pair is predicted in K562 cell but not GM12878 cell.

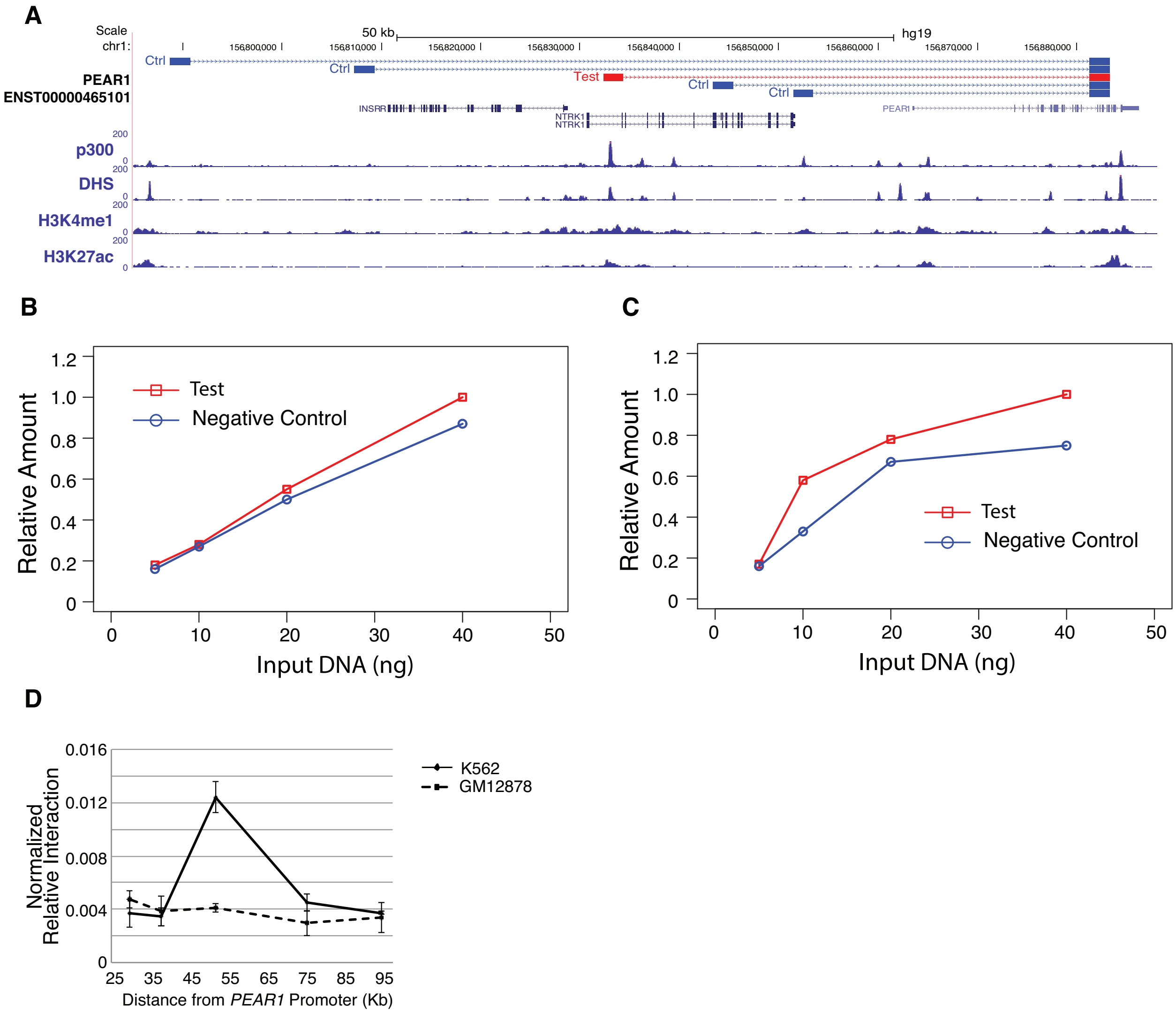


Fig. S13. Cumulative distributions of enhancers and EP pairs with decreasing degree of cell-type specificity. Enhancers are predicted with 1% FDR cutoff. Given this set of enhancers, EP pairs were predicted at varied FDR cutoffs. P-values are for comparing EP pair and enhancer curves using KS test.

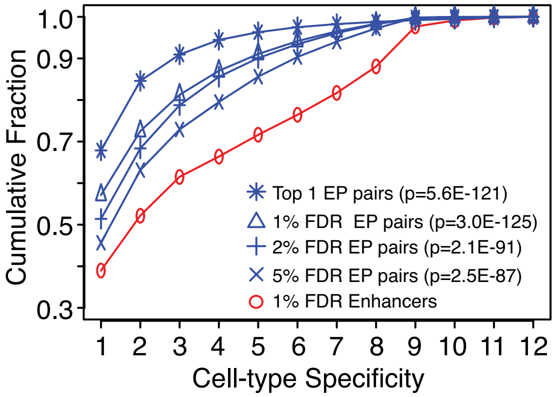


Fig. S14. Example shadow enhancers. (A) K562 cell; **(B)** MCF-7 cell. Interactions detected by ChIA-PET are depicted by pairs of boxes connected by a thin line. Boxes with darker shade indicate interactions with higher confidence.

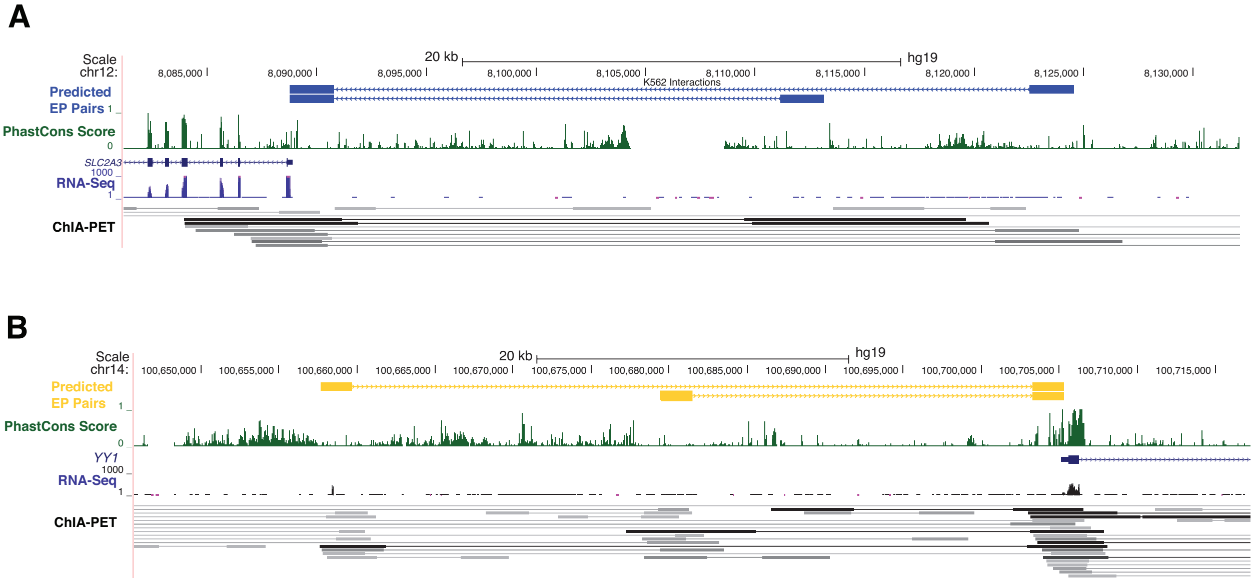


Fig. S15. Gene expression specificity of transcription factors with binding sites at mirrored CNC and CAC sites. P-value is based on one-sided Student's t-test. N(CNC) = 60,306, N(CAC) = 13,320.

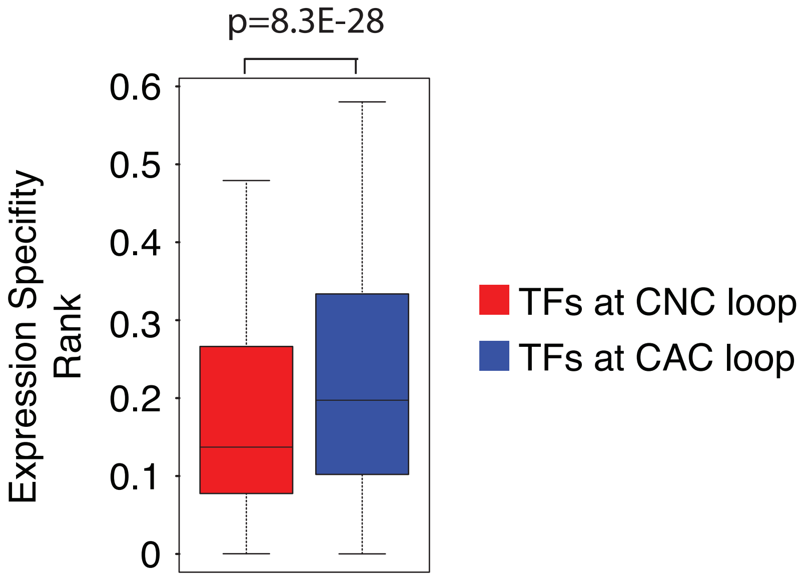
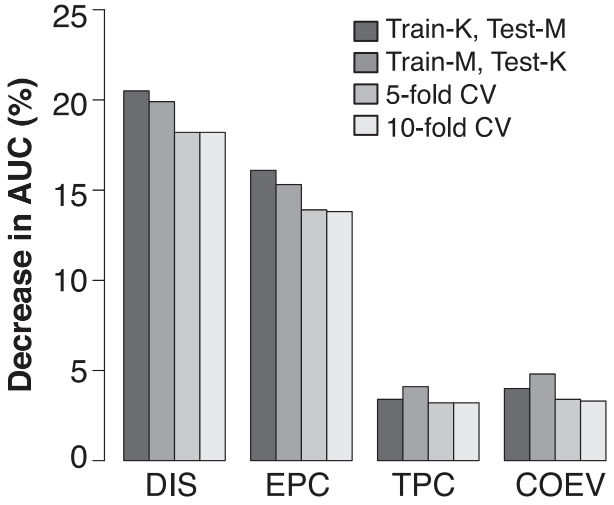


Fig. S16. Relative importance of features used in the IM-PET model. Features were removed one at a time. Random Forest model was trained with the rest of the features and evaluated using different cross validation schemes. Train-K, Test-M: train with K562 positive and negative EP pairs and test MCF-7 positive and negative pairs. Train-M, Test-K, vice versa. AUC, area under the curve.



Supplementary Tables

Table S1. Overlap of predicted enhancers with other genomic marks for enhancers. NA, ChIP-Seq data not available.

Cell Type	# of Enhancers	# DHS (%)	# Conserved (%)	# p300 (%)	# with at least one (%)
GM12878	12696	8530 (67)	4945 (39)	8232 (65)	11233(88)
H1ESC	13906	11843 (85)	6755 (49)	7088 (51)	13066(93)
HepG2	19648	14770 (75)	7235 (37)	13907 (71)	17825(90)
HMEC	18689	11784 (63)	7951 (43)	NA	14580(78)
HSMM	14049	12206 (87)	6258 (45)	NA	12990(92)
HUVEC	19445	15404 (80)	8611 (44)	NA	17288(87)
IMR90	16825	13292 (79)	7908 (47)	9590 (57)	15479 (92)
K562	15987	13720 (86)	6214 (39)	6606 (41)	14945(93)
MCF-7	16940	9726 (57)	5744 (34)	1427 (8)	12127(72)
NHEK	19111	15131 (79)	7919 (41)	NA	16700(87)
NHLF	15650	11843 (76)	6967 (45)	NA	14110(90)
CD4$^+$ T	21796	10983 (50)	7401 (34)	1818 (8)	14703(67)

Table S2. Performance of Random Forest models trained using different RF parameters and evaluated using different cross validation schemes. Train-K, Test-M: train with K562 positive and negative EP pairs and test MCF-7 positive and negative pairs. Train-M, Test-K, vice versa. AUC, area under the curve.

Num of trees in model	Num of features in each tree	Maximum depth of trees	AUC			
			Train-K, Test-M	Train-M, Test-K	5-fold Cross Validation	10-fold Cross Validation
1000	2	∞	0.92	0.90	0.94	0.94
500	2	∞	0.92	0.90	0.94	0.94
200	2	∞	0.92	0.90	0.93	0.94
100	2	∞	0.92	0.90	0.93	0.94
50	2	∞	0.92	0.90	0.93	0.94
10	2	∞	0.90	0.88	0.93	0.92
1000	1	∞	0.92	0.90	0.94	0.94
1000	3	∞	0.92	0.90	0.93	0.93
1000	2	50	0.92	0.90	0.94	0.94
1000	2	10	0.92	0.90	0.92	0.92

Table S3. List of qPCR primers used in the 3C-qPCR experiments. Genomic coordinates for the center of enhancer and transcription start site are shown after the name of each test interaction, which is in the format of "cell type-target gene".

Primer Type	Sequence	Start	End
GM12878-DDX39B (chr6: 31462622- 31509758)			
Bait	CCTGCTCAAAGAATCTGGTTAGT	31511074	31511096
Test	GGGCAACAAGAGCGAAACT	31462581	31462600
Negative control 1	AGGAAAGCACAGTGGAAGGT	31441775	31441795
Negative control 2	GTGGGACTCTCTGTCATCTTCA	31448998	31449019
Negative control 3	GGATGGTCTCTTCACTTTGTTTA	31469157	31469180
Negative control 4	ATTCTCTAAGTGAATCATGTAACCA	31484531	31484556
GM12878-HLA-DQA1 (chr6: 32633823-32605208)			
Bait	GTGTCACCTCACAAGTAATCAAAT	32605906	32605929
Test	TGGAAAGGACCTACACCTCTGA	32633379	32633400
Negative control	ATACTCTACAAACACAAGCAACCA	32662529	32662552
GM12878-BATF (chr14: 76009248-75988784)			
Bait	GGGCAGCGAACACTGATAGA	75986985	75987004
Test	CGTACAGGGGCTGGTAACTG	76006267	76006286
Negative control	TCCTCATTTCCATCTGACACCT	76060503	76060524
GM12878-CD53 (chr1: 112136278-111440292)			
Bait	ACATTGATGTCCTCACTAAGAAAA	111433293	111433316
Test	TACTAACTGCTGAACATCCCTCT	112131281	112131303
Negative control 1	CAGATTTGGCTCAGGAGTCATA	112083793	112083816

Negative control 2	AAGGTGAACACTCAGAACAAAGA	112101159	112101182
Negative control 3	TTCCTGAGTGAAGGGATGGT	112156935	112156955
Negative control 4	GATAGAACCCTTAGTAAATGACCAG	112188814	112188838
GM12878-POU2AF1 (chr11: 111287791-111250417)			
Bait	TGCCCACCCACTGATAACA	111250747	111250766
Test	CTACAGCCAATCAGTTCAGGA	111290009	111290029
Negative control 1	CGGTTGAAACTGGAGTGGTA	111262606	111262626
Negative control 2	CAAAACTGACCCTCTTTATCGT	111276220	111276242
Negative control 3	CCCTTTTCAGATTTTTGTTCAC	111311377	111311398
Negative control 4	TAAGTCTCAGCAACGAATGGTA	111349354	111349376
K562-MYCL3 (chr8: 129569619-128748477)			
Bait	CCCCAATAAATCCAGTGTCTT	128745915	128745935
Test	GCAGAAAATAAATTGTCCAAGTT	129568748	129568770
Negative control	TGCTGAATACTTGAGGTTAGACTT	129149376	129149399
K562-DDX39B (chr6: 31509758-31462422)			
Bait	CCTGCTCAAAGAATCTGGTTAGT	31511074	31511096
Test	GGGCAACAAGAGCGAAACT	31462581	31462600
Negative control 1	AGGAAAGCACAGTGGAAGGT	31441775	31441795
Negative control 2	GTGGGACTCTCTGTCATCTTCA	31448998	31449019
Negative control 3	GGATGGTCTCTTCACTTTGTTTA	31469157	31469180
Negative control 4	ATTCTCTAAGTGAATCATGTAACCA	31484531	31484556
K562-GTSF1 (chr12: 53057134- 54867386)			
Bait	CATCCCAATCTTCAGTGCTAA	54866105	54866125
Test	CTCCTCATCACTCTCCCCAG	53054077	53054096
Negative control 1	CACTTCTTCTCTTTCACGGACT	52985986	52986008
Negative control 2	CCTCACCACCCTACCTCACT	53010988	53011007
Negative control 3	ACAGGTGGTAGAAACAAGAGCA	53099729	53099751
Negative control 4	AGTTGTGGGATTCCTGCCT	53168272	53168291
K562-PEAR1 (chr1: 156833377- 156883324)			
Bait	CTGGAAATAATCATTTGTGAGTCA	156881874	156881897
Test	CGCTGCTTGTTTGCTGGT	156831007	156831026
Negative control 1	GGATTGTCTGTTTACTCTGCTGA	156789709	156789731
Negative control 2	GTGTTCATCCTTCCTTCTCCA	156808276	156808297
Negative control 3	GTGGAGAAGAAGGACGAAACA	156844389	156844410
Negative control 4	CCCTATCACTTCCAATCACCT	156852439	156852460
Internal control primers			
GAPDH-F	ATGTTCGTCATGGGTGTGAA	6646327	6646346
GAPDH-R	AGGCATTGCTGCAAAGAAAG	6646463	6646482

Table S4. List of BAC clones used in 3C-qPCR experiments.

BAC ID	ID of test interaction covered
RP11-184F16	GM12878-DDX39B
RP11-257P24	GM12878-HLA-DQA1
RP11-17G1	GM12878-BATF
RP11-705K13	GM12878-CD53

RP11-631D1	GM12878-CD53
RP11-878N13	GM12878-POU2AF1
RP11-243J12	K562-MYCL3
RP11-440N18	K562-MYCL3
RP11-184F16	K562-DDX39B
RP11-441M5	K562-GTSF1
RP11-753H16	K562-GTSF1
RP11-730I22	K562-PEAR1

Table S5. Gene ontology term enrichment analysis of genes regulated by shadow enhancers. Top five enriched GO biological process terms are shown. Adjusted (Benjamini-Hochberg procedure) p-value cutoff is 0.05**.** Cell-type-specific GO terms are highlighted in blue.

Cell Type	**Enriched in genes with 3 or more enhancers**	**Enriched in genes with 2 or fewer enhancers**
GM12878 (lymphoblastoid cell)	• Viral reproduction • Regulation of innate immune response • Immune response – acting signal transduction • regulation of I-kappaB kinase/NF-kappaB cascade • mature B cell differentiation	• mRNA metabolic process • Macromolecule metabolic process • Intracellular transport • Cell cycle phase • Protein localization
H1-ESC (Embryonic Stem Cell)	• Regulation of telomere maintenance • Chordate embryonic development • Erythrocyte differentiation • Regulation of cyclin-dependent protein kinase activity • Embryonic morphogenesis	• Regulation of transcription, DNA-dependent • Nervous system development • Cellular protein metabolic process • RNA biosynthetic process • Cell part morphogenesis
HepG2 (hepatocellular carcinoma)	• Response to DNA damage • Cellular response to stress • Glycerolipid biosynthetic process • Cholesterol metabolic process • Insulin receptor signaling pathway	• Cell cycle phase • Cellular protein localization • mRNA metabolic process • Intracellular transport • Chromatin organization
HMEC (mammary epithelial cell)	• Epithelium development • Mammary gland morphogenesis • Epithelial cell differentiation • Keratinocyte differentiation • Regulation of epithelial cell migration	• Intracellular transport • Nitrogen compound biosynthetic process • Protein catabolic process • RNA processing • Cell cycle process
HSMM (skeletal muscle myoblast)	• Actin cytoskeleton organization • Regulation of epithelial cell proliferation • Adherens junction organization	• Cellular protein metabolic process • RNA process • Nitrogen compound

	• Cardiac muscle tissue growth • Fibroblast growth factor receptor signaling pathway	biosynthetic process • Apoptosis • Protein transport
HUVEC (umbilical vein endothelial cell)	• Vasculature development • Regulation of epidermal growth factor receptor signaling pathway • Heart morphogenesis • Muscle cell development • Placenta development	• Protein ubiquitinatioin • Intracellular protein transport • Cellular macromolecule catabolic process • RNA biosynthetic process • Cellular metabolic process
IMR90 (fetal lung fibroblast)	• Lung development • Tube development • Blood vessel development • Morphogenesis of an epithelium • Muscle cell differentiation	• Nucleic acid metabolic process • Protein metabolic process • Regulation of gene expression • Regulation of cell motion • Establishment of protein localization
K562 (myelogenous leukemia)	• DNA damage response • Endosome transport • Cell development • Regulation of endocytosis • Regulation of innate immune response	• Establishment of protein localization • Cellular response to stress • RNA metabolic process • Regulation of transcription, DNA-dependent • Nitrogen compound biosynthetic process
MCF7 (breast ductal carcinoma)	• Embryonic development ending in birth or egg hatching • Heart development • Tube morphogenesis • Embryonic appendage morphogenesis • Gland morphogenesis	• Regulation of transcription, DNA-dependent • Biopolymer modification • Virus-host interaction • RNA metabolic process • Intracellular transport
NHEK (epidermal keratinocyte)	• Regulation of epidermal growth factor receptor signaling pathway • Epithelial cell differentiation • Cell-cell junction assembly • Keratinocyte differentiation • Cellular response to extracellular stimulus	• Nitrogen compound biosynthetic process • Protein localization • Cell cycle phase • Regulation of cell development • Cell death
NHLF (lung fibroblast)	• Fibroblast growth factor receptor signaling pathway • Ameboidal cell migration • Regulation of epithelial cell proliferation • Regulation of fibroblast proliferation • Lung cell differentiation	• Cellular protein localization • Virus-host interaction • Intracellular transport • Cell cycle phase • Protein catabolic process
CD4+ T (peripheral blood T cell)	• T cell differentiation • Lymphocyte activation • T cell mediated immunity • Immune system development • T-helper cell differentiation	• Establishment of protein localization • Cell cycle process • Intracellular transport • RNA processing

		- Cellular protein metabolic process

Table S6. Overlap of EP pairs with mirrored CNC and CAC sites. P-values are based on Fisher's exact test.

Cell Type	Cohesin Subunit	# of EP pairs overlap with CNC	p-value	# of EP pairs overlap with CAC	p-value
GM12878	Rad21, Smc3	5047	6.1E-22	604	0.65
HepG2	Rad21, Smc3	4691	6.9E-44	665	1.8E-2
K562	Rad21, Smc3	4693	1.6E-9	991	1.7E-2
H1ESC	Rad21	263	3.1E-7	1080	1.5E-2
MCF-7	Rad21, Stag1	3143	9.0E-3	260	0.51

References

1. Li G*, et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148(1-2):84-98.
2. Firpi HA, Ucar D, & Tan K (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* 26(13):1579-1586.
3. Calo E & Wysocka J (2013) Modification of enhancer chromatin: what, how, and why? *Molecular cell* 49(5):825-837.
4. Li G*, et al.* (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology* 11(2):R22.
5. Dekker J, Rippe K, Dekker M, & Kleckner N (2002) Capturing chromosome conformation. *Science* 295(5558):1306-1311.
6. Wang Z*, et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics* 40(7):897-903.
7. Chepelev I, Wei G, Tang Q, & Zhao K (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic acids research* 37(16):e106.
8. Harrow J*, et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* 22(9):1760-1774.
9. Bryne JC*, et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* 36(Database issue):D102-106.
10. Matys V*, et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* 34(Database issue):D108-110.
11. Newburger DE & Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research* 37(Database issue):D77-82.
12. Ernst J*, et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43-49.
13. Hon GC, Hawkins RD, & Ren B (2009) Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* 18(R2):R195-201.

14. Won KJ, Chepelev I, Ren B, & Wang W (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC bioinformatics* 9:547.
15. Schug J*, et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* 6(4):R33.
16. Ucar D, Hu Q, & Tan K (2011) Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic acids research*.
17. Breiman L (2001) Random Forests. *Machine Learning* 45(1):5-32.
18. Schadt EE*, et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS biology* 6(5):e107.
19. Innocenti F*, et al.* (2011) Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS genetics* 7(5):e1002078.
20. Stranger BE*, et al.* (2007) Population genomics of human gene expression. *Nature genetics* 39(10):1217-1224.
21. Veyrieras JB*, et al.* (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics* 4(10):e1000214.
22. Degner JF*, et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482(7385):390-394.
23. Gaffney DJ*, et al.* (2012) Dissecting the regulatory architecture of gene expression QTLs. *Genome biology* 13(1):R7.
24. Pickrell JK*, et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289):768-772.
25. Montgomery SB*, et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464(7289):773-777.
26. Dimas AS*, et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325(5945):1246-1250.
27. Chepelev I, Wei G, Wangsa D, Tang Q, & Zhao K (2012) Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell research* 22(3):490-503.
28. Jin F*, et al.* (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503(7475):290-294.
29. Thurman RE*, et al.* (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75-82.