Supplementary Information
A framework for supervised enhancer prediction with epigenetic pattern recognition and targeted validation across organisms

Methods

Creation of Metaprofile:

We utilized the smoothed histone signal tracks provided for the S2 cell-line by the modENCODE consortium [1] to aggregate the corresponding histone signals around the STARR-seq peaks [2]. This aggregation was performed to remove noise before using the metaprofile *s(n)* for identifying active regulatory regions in the genome. The genome-wide profile for open chromatin (DNase-seq or DHS) for the S2 cell-line was calculated based on the experiments by the Stark lab [2]. To create the smoothened metaprofile, we aggregated the H3K27ac signal of active STARR-seq peaks with a noticeable “double peak” pattern within the H3K27ac signal in the S2 cell-line. All the STARR-seq peaks that overlap with DHS or H3K27ac peaks are assumed to be active regulatory regions in the genome.

To identify double peak regions, we initially identified the minimum in the H3K27ac signal track closest to the middle of the STARR-seq peaks. A minimum is accepted if it has the lowest signal within a 100 base pair region in the H3K27ac signal track. Then we proceed to identify the flanking maxima (both sides of the minimum) within a total of 2-kilo base pair region of the STARR-seq peak (1kb on each direction from the center of the STARR-seq peak). These maxima are accepted only if they have the highest signal within a 100 base pair region in the H3K27ac signal track. Approximately 70% of the active STARR-seq peaks contained an identifiable double peak within the H3K27ac signal.

After identifying the double peaks surrounding STARR-seq peaks, we aggregated the signal after aligning the maxima flanking the regulatory region. The signal track is interpolated with a cubic spline fit so that the signal track contains equal number of points for each double peak region. All interpolation and smoothing steps were performed using the scipy module in python. The aggregated signal tracks are averaged to create the metaprofile for the double peak regions. While the signal tracks are aggregated based on identifying the double peak regions in the H3K27ac signal track, the same set of operations can be performed with any epigenetic mark expected to have the double peak pattern flanking regulatory regions.

In addition, while creating the metaprofile for H3K27ac signal close to active STARR-seq peaks, we also performed the same set of transformations on other dependent epigenomic datasets (other histone marks and/or DHS signal). In this study (Figures 1 and S2), the dependent profiles for all other epigenetic datasets are calculated by averaging the corresponding signal based on identifying double peak regions within H3K27ac signal. If the signal tracks of the other epigenetic marks also tend to contain a double peak pattern in the same regions, the metaprofiles for the corresponding epigenetic marks will also contain a double peak pattern as observed in Figure S2A. However, as DHS and repressive histone marks do not contain a double peak pattern

(Figure S2), these regions do not have the same epigenetic template associated with enhancers.

Matched Filter Algorithm:

The epigenetic signal at enhancers and promoters can be approximated as the linear superposition of background noise and the metaprofile *s(n)* learned in Figure 1 (Figure S2) for the corresponding experimental dataset. The matched filter *h(n)* is used to scan the epigenetic signal to identify the occurrence of the metaprofile pattern within different regions of the genome. Before calculating the matched filter score, interpolation of signal is used to ensure that the scanned region contains the same number of points as the metaprofile. The matched filter process is equivalent to the computation of the cross correlation between the signal *y(n)* and the reverse of the transformed metaprofile template *s*(N-n)* (where *N* is the total number of points in the template). In other words:

$$r(n) = \sum_{i=1}^{N} y(i) * h(i)$$

where *h(i)* is the matched filter and can be written as:

$$h(i) = s^{*}(N - i)$$

As shown in Figure S1, there is a large amount of variability in the span (distance between the two peaks in the histone signal) of the regulatory region in the epigenetic signal. As a result, we scan the genome with the matched filter scanning different spans of the genome (distance between the two peaks allowed to vary between 300 and 1100 base pairs) and take the highest score as the matched filter score for that region. The matched filter is the filter that recognizes any given template in the presence of noise in a signal with the highest signal-to-noise ratio [3]. In the presence of white noise alone, the matched filter score is low and follows a Gaussian distribution (negatives). The presence of the metaprofile within the signal leads to higher matched filter scores for positives.

Statistical Learning Models

The matched filter scores for negatives for different histone marks are unimodal that can be fit using separate Gaussian distributions. The Z-scores of matched filter scores with respect to the negatives (random regions of genome) are used as input features for training different statistical learning models. The Z-score of the matched filter score for a region (*z(i)*) is:

$$z(i) = \frac{r(i) - \mu}{\sigma}$$

where *r(i)* is the matched filter score for region *i* while μ *and* σ are the mean and standard deviation of the Gaussian fit to the matched filter scores for random regions in genome. In the main text, we discuss our results of the Support Vector Machine (SVM) model, which is one of the most versatile and successful binary classifiers [4]. We utilized a linear kernel to distinguish between the positives and negatives. The linear SVM identifies a decision boundary that maximally discriminates the epigenetic features of regulatory regions from random regions of the genome in the SVM feature vector space.

In Figure S5, we also present results for Ridge Regression [5], Random Forest [6], and Gaussian Naïve Bayes [7] models and the accuracy of different models are comparable. Ridge regression is a linear regression technique that prevents over fitting by penalizing large weights for each feature. Random Forest is an ensemble learning method that operates by constructing a large number of decision trees and outputting the mean prediction of different decision trees. We used thousand trees for creating our enhancer and promoter prediction models. The naïve Bayes classifier is a family of simple probabilistic classifiers that assumes that all the features are independent of one another. We used scikit-learn [8] with default parameters for training and assessing the performance of all the statistical models. In general, the SVM and random forest models performed the best over all the tests and were the most flexible models.

Model Assessment

In order to assess the accuracy of matched filter for predicting enhancers and promoters, we used 10-fold cross validation. During 10-fold cross validation, the positives and negatives are randomly divided in to 10 groups each. Nine of the 10 groups are randomly combined to train the model and the predictions are tested on the 10th group. To evaluate the performance of trained classifiers, we performed 10-fold cross-validation on the training data and quantified our results with area under receiver-operating characteristic (ROC), and area under precision-recall (PR) curves.

In the ROC curve [9], the true positive (TP) rate is plotted against the false positive (FP) rate at different thresholds in the statistical model. The TP rate is defined as the fraction of positives identified correctly by the model (i.e., ratio of number of true positives identified by the model to the total number of positives). The FP rate is defined as the fraction of negatives identified correctly by the model (i.e., ratio of number of negatives misclassified by the model to the total number of negatives). While comparing the performance of two different classifiers in the ROC curve, the classifier with higher TP rate at the same FP rate is considered to be a better classifier. The area under the ROC is a single measure for the accuracy of a model as models with higher area under ROC are generally considered to be better models.

In the PR curve, the precision is plotted against recall at different thresholds in the statistical model. The recall is the same as the TP rate of the model (i.e., ratio of number of true positives identified by the model to the total number of real positives). The precision is the fraction of positives in the model that are correct (i.e., ratio of number of true positives identified by the model to the total number of positives according to the model). In skewed datasets with large number of negatives in comparison to positives, the FP rate can be low even when the number of false positives misclassified by the model is comparable to the number of true positives. For such skewed datasets, the area under ROC for two different models may be very similar even though they actually differ in performance with respect to their precision. Hence, the area under the PR curve is a better reflection of the performance difference between two models with similar area under ROC in skewed datasets.

In Figure 2, the positives are defined as the active peaks (intersecting with DHS or H3K27ac peaks) from a single STARR-seq experiment (singe core promoter) or the union of active peaks from multiple STARR-seq experiments (multiple core promoters). The negatives are randomly chosen regions in the genome with H3K27ac signal that had the same width distribution as the distribution of distance between double peaks near STARR-seq peaks (shown in Figure S1). We typically chose between 5 to 10x number of negatives as compared to number of positives in Figures 2, 3, and 4 as the number of enhancers and promoters in the genome (positives) are far lesser than the number of negatives and area under PR curve is dependent on the ratio of negatives to positives during 10-fold cross validation. The matched filter score for each region is chosen as the best matched filter score with a 1500 bp region centered on each positive and negative. The matched filters are scanned with distances between 300-1100 bp before choosing the best score. While comparing the performance of the matched filter to the peak-based models of the different epigenetic marks (Figure S4), we assumed that histone (DHS) peaks that overlapped with at least 50% (10%) of the STARR-seq peak is used to rank that prediction. We used a smaller threshold for DHS peaks as they are much smaller than histone peaks. We achieved similar results with thresholds of 25% for both histone and DHS peaks. The p-value of the intersecting peak is used to rank the peak-based predictions. The modENCODE histone peaks [1] and DHS peaks [2] were compared to the matched filter scores in Figure S4.

During STARR-seq, each peak is functioning as an enhancer within the plasmid environment in S2 cell-line. However, to delineate the native role of the region, we classify them as promoters and enhancers based on their distance to the transcription start sites in the genome. In Figure 3, the active promoters are defined as active STARR-seq peaks (multiple core promoter) within 1 kb of TSS (Ensembl release 78) while enhancers were active STARR-seq peaks more than 1kb from any TSS in *Drosophila melanogaster*. While calculating the matched filter for positives and negatives, we considered the best scoring matched filter score after padding each region to 1.5kb width.

Transgenic mouse enhancer assay

In Figure 4, the enhancers were tested in transgenic mouse reporter assay [10,11]. Predicted enhancers were PCR amplified and cloned into a plasmid upstream of a minimal hsp68 promoter and *lacZ* reporter gene. Resulting plasmids were linearized and injected into single-cell FVB/NCrl strain *Mus musculus* embryos. After reimplantation into surrogate mothers, resulting embryos were collected at embryonic day 11.5 (E11.5), stained for b-galactosidase activity, and imaged. Elements were scored positive for enhancer activity if at least three resulting transgenic embryos had reporter gene expression in the same tissue and pattern. Elements were scored negative if at least five transgenic embryos were recovered and no reproducible staining patterns was observed. Enhancer names (mm numbers) reported here are the unique identifiers from the VISTA Enhancer Browser (www.enhancer.lbl.gov).

All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare Committee. All mice used in this study were housed at the Animal Care Facility (the ACF) at LBNL. Mice were monitored daily for food and water intake, and animals were inspected weekly by the Chair of the Animal Welfare and Research Committee and the head of the animal facility in consultation with the veterinary staff. The LBNL ACF is accredited by the American Association for the Accreditation of Laboratory Animal Care International (AAALAC)

Assessment with mESC FIREWACh assay peaks

In Figure S12, the promoters are defined as FIREWACh peaks within 2 kb of TSS (GENCODE release vM4) while enhancers were FIREWACh peaks more than 2kb from any TSS. The larger distance (2 kb) for defining promoters was used because of the larger size of the mouse genome. The FIREWACh assay is performed in a transduction assay and was based on ChIP-seq peaks of a few key TFs. Hence, we did not split the FIREWACh peaks in to active and poised enhancers and promoters. The ENCODE histone and DHS datasets for mESC were used to predict enhancers and promoters in Figure S12.

H1-hESC whole genome prediction

To predict enhancers and promoters on the whole genome, we utilized the 6 parameter machine learning model shown in Figure 2. The histone and DHS signals from ENCODE consortium [12] were used to predict enhancers and promoters in H1-hESC. The histone signals were converted to log fold enrichment (with respect to control signal) before we scanned it with the matched filter. There were 43463 active regulatory regions predicted in the human genome (< 2% of genome). All regions within 2kb of TSS were annotated as promoters while active regulatory regions that were more than 2kb from TSS were annotated as enhancers. The distribution of the expression of closest gene (GENCODE v19 TSS) from ENCODE RNA-seq dataset [12] for H1-hESC was compared to the expression of all genes from H1-hESC. The Wilcoxon test was used to measure the significance of changes in gene expression.

Overlap with chromatin state predicted by chromHMM and SegWay

We compared the promoter and enhancer predictions for the H1-hESC cell-line with the chromatin states for the H1-hESC cell-line predicted by chromHMM and SegWay. The chromatin states for H1-hESC were downloaded from the ENCODE portal. The prediction is considered to be overlapping with the corresponding chromatin state if more than 50% of the predicted enhancer or promoter is labeled as the same chromatin state.

Enhancer Validation Experiment

Cell lines

WA01 or H1 hESC was obtained from WiCell and maintained feeder-free on matrigel-coated plates in mTESR1medium (StemCell Technologies) supplemented with penicillin and streptomycin. Roughly once weekly cell colonies were dissociated using dispase and absence of differentiation was confirmed by visual inspection and periodically staining cells using anti-SSEA4 conjugated to FITC and performing flow cytometry. Other cell types (HOS and A549 obtained from ATCC and TZMbl from the AIDS Reagent Repository) were maintained in DMEM supplemented with 10% fetal calf serum and passaged twice weekly using trypsin.

Preparation of HIV vector, cellular transduction and analysis

Self-inactivating (SIN) HIV vector pFG12 was modified in that the UBC promoter driving eGFP along with the WPRE was removed and replaced with a 1.4 kb IRES-eGFP cassette. Upstream of the IRES a 142 bp basal Oct 4 promoter (5'-CCTCCCTCTCCTCCACCCATCCAGGGGGCGGGGCCAGAGGTCAAGGCTAGTGGGTGGGACTGGGGAGGGAGAGAGGGGTTGAGTAGTCCCTTCGCAAGCCCTCATTTCACCAGGCCCCCGGCTTGGGGCGCCTTCCTTCCCC-3'; coordinates on chromosome 6, negative strand: 31138398-31138539) was inserted, which overlaps with the TSS of *Oct4* but not with the coding sequence. A unique Xba 1 site was present just upstream of the basal Oct4 promoter, for cloning of test insert DNA fragments. Each test DNA fragment was amplified from genomic DNA using nested PCR and Takara LA enzyme. Typical initial PCR amplification conditions were 98^{o}C for 10 s, 55^{o}C for 15 s, and 68^{o}C for 3 min for 30 cycles using 100-200 ng of genomic DNA, with the annealing temperature being variable depending upon the T_m of the primer pair. For the second (internal) round of PCR, only 1-2% of the original product was used under similar PCR conditions, but for 15 cycles.

PCR products were individually cloned into TOPO pCRII-blunt vector (Invitrogen) and insert identity confirmed by both restriction digests and dideoxy sequencing. All DNA inserts were cloned into the unique Xba 1 site of the HIV vector described above using compatible cohesive ends, and in each case both orientations of the insert within the vector were confirmed by appropriate restriction digests.

HIV vector supernatants were prepared by co-transfecting 35 mm tissue culture wells of 293T cells (~75-80% confluence), each with 5 µg of HIV transfer vector (HIV-TV) with DNA element of interest, HIV packaging vector, and pME VSV G (encoding Indiana strain VSV G). After 48-72 hours, vector supernatant was harvested, centrifuged at 3000 x *g* for 10 min, and stored at -80^o C until use.

In order to transduce the WA01 hESC, cells were first lifted using dispase, washed extensively, and plated in the presence of ROCK Inhibitor Y-27632 (StemCell Technologies) on matrigel-coated plates. After a few hours, cells were transduced for 4-6 h with lentiviral vector supernatant, After 48-72 h single cell suspensions were again prepared using dispase and Y-27632 and cells were analyzed for eGFP expression as described above, collecting 10,000 events. For all other cell lines, cells were plated the day before in 12 well format, transduced using the indicated amounts of vector supernatant, refed the following day, and analyzed for eGFP expression 48-72 h later, as described above.

The fold change of inactive elements was used to calculate the background distribution of inactive elements. This was fit to a normal distribution and putative enhancers that displayed higher activity than expected by chance (p-value < 0.05) were considered to be active in the cell-line. This was done for the forward and reverse directions separately and elements that were positive in either orientation were considered to be active.

H1-hESC TF binding

To measure the differences in TF binding and co-binding patterns at promoters and enhancers, we overlapped the ChIP-seq peaks from ENCODE with our predicted enhancers and promoters using intersectBed. The two regions were considered to be overlapping if at least 25% of the ChIP-seq peak was overlapping with the predicted enhancer or promoter.

Table S1 – Performance of matched filter models with single epigenetic feature for all STARR-seq peaks (multiple core promoters)

Feature	AUROC	AUPR
H3K27ac	0.95	0.90
H3K4me1	0.70	0.59
H3K4me2	0.91	0.79
H3K4me3	0.84	0.76
H3K9ac	0.92	0.85
H4K12ac	0.92	0.86
H3	0.80	0.70
H1	0.88	0.81
H2BK5ac	0.94	0.90
H4K8ac	0.88	0.79
H4K5ac	0.87	0.79
H4K16ac	0.89	0.72
H3K18ac	0.90	0.84
H3K9me1	0.71	0.61
H3K79me2	0.79	0.58
H4K27me2	0.81	0.68
H2Av	0.66	0.57
H3K27me3	0.83	0.64
H3K23ac	0.66	0.46
H3K79me3	0.70	0.51
H3K27me1	0.64	0.43
H4	0.67	0.49
H3K36me1	0.54	0.41
H3K9me3	0.59	0.42
H3K9me2	0.60	0.41
H3K36me3	0.57	0.38
H4K20me1	0.47	0.31
H3K79me1	0.47	0.30

Table S2 – Performance of matched filter models with single epigenetic feature for promoters and enhancers (multiple core promoters). Numbers within (outside) parenthesis are accuracy of models for predicting promoters (enhancers).

Feature	AUROC	AUPR
H3K27ac	0.91 (0.96)	0.60 (0.73)
H3K4me1	0.88 (0.60)	0.42 (0.16)
H3K4me2	0.84 (0.92)	0.21 (0.48)
H3K4me3	0.62 (0.92)	0.09 (0.65)
H3K9ac	0.85 (0.94)	0.24 (0.70)
H4K12ac	0.90 (0.93)	0.33 (0.58)
H3	0.78 (0.83)	0.26 (0.48)
H1	0.83 (0.92)	0.36 (0.61)
H2BK5ac	0.91 (0.96)	0.59 (0.70)
H4K8ac	0.90 (0.86)	0.55 (0.37)
H4K5ac	0.89 (0.86)	0.52 (0.41)
H4K16ac	0.90 (0.90)	0.52 (0.40)
H3K18ac	0.90 (0.88)	0.60 (0.47)
H3K9me1	0.53 (0.81)	0.09 (0.44)
H3K79me2	0.70 (0.83)	0.10 (0.27)
H4K27me2	0.68 (0.85)	0.19 (0.44)
H2Av	0.63 (0.78)	0.15 (0.36)
H3K27me3	0.81 (0.86)	0.20 (0.36)
H3K23ac	0.55 (0.71)	0.07 (0.20)
H3K79me3	0.61 (0.74)	0.08 (0.23)
H3K27me1	0.72 (0.57)	0.12 (0.12)
H4	0.69 (0.68)	0.13 (0.21)
H3K36me1	0.75 (0.58)	0.19 (0.18)
H3K9me3	0.59 (0.64)	0.11 (0.15)
H3K9me2	0.62 (0.63)	0.09 (0.15)
H3K36me3	0.60 (0.62)	0.09 (0.14)
H4K20me1	0.55 (0.50)	0.07 (0.10)
H3K79me1	0.54 (0.58)	0.06 (0.12)

Table S3 Summary of predicted mouse regulatory regions in six different tissues

Tissue	Regulatory regions	Distal regulatory regions	Proximal regulatory regions
Forebrain	35509	24423 (68.8%)	11086 (31.2%)
Hindbrain	32855	22659 (69.0%)	10196 (31.0%)
Limb	38232	26761 (70.0%)	11471 (30.0%)
Midbrain	33451	22947 (68.6%)	10504 (31.4%)
Heart	30739	20282 (66.0%)	10457 (34.0%)
Neural Tube	38933	27033 (69.4%)	11900 (30.6%)

Table S4 Transgenic mouse reporter assays results for 31 elements in E11.5

element #	Name	hg19 coordinates	Result summary
2346	EN202	chr4:23932061-23933692	**8/11 eye, 5/11 facial mesenchyme**
2349	EN205	chr22:47048605-47050100	Negative
2353	EN209	chr10:97267716-97269342	**4/6 heart**
2357	EN214	chr1:214280595-214282080	**8/11 heart**
2359	EN216	chr3:42113230-42114717	Negative
2371	EN228	chr17:55618678-55620173	Negative
2372	EN229	chr2:109252387-109254056	Negative
2373	EN230	chr20:43201171-43202669	Negative
2374	EN231	chr1:225954390-225955885	**4/5 branchial arch**
2375	EN232	chr17:71287045-71288497	Negative
2377	EN234	chr6:163630391-163631925	Negative
2378	EN235	chr11:12203825-12205249	Negative
2380	EN237	chr20:46012576-46013656	Negative
2382	EN240	chr3:186123841-186125332	Negative
2384	EN242	chr2:20778294-20779806	**10/10 heart, 7/10 ear, 5/10 other**
2387	EN245	chr7:130012949-130014460	Negative
2393	EN251	chr20:17839843-17841338	Negative
2394	EN252	chr6:108909808-108911282	Negative
2397	EN255	chr6:46020500-46022001	Negative
2399	EN257	chr6:43760764-43762277	Negative
2400	EN258	chr21:29655315-29656764	Negative
2403	EN261	chr11:8753701-8755208	Negative
2404	EN262	chr1:203660971-203662806	Negative
2405	EN263	chr6:17931980-17933492	Negative
2412	EN270	chr4:129278773-129280245	Negative
2414	EN272	chr4:47826466-47828052	**5/5 heart**
2415	EN273	chr22:28028233-28029715	Negative
2417	EN275	chr4:128406285-128407745	Negative
2418	EN276	chr1:92310736-92312231	Negative
2419	EN277	chr7:82039621-82041108	**12/12 somites; 11/12 limb, 10/12 eye, 9/12 brachial arch**
2420	EN278	chr10:5627988-5629809	Negative

Table S5 Transgenic mouse reporter assays results for 102 elements in E11.5

element #	Name	mm9 coordinates	Result summary
1303	mEN351	chr10:61532677-61537653	Negative
1304	mEN352	chr15:75646709-75649708	**4/7 forebrain**
1305	mEN353	chr9:121301588-121305883	Negative
1332	mEN354	chr4:135257075-135260072	Negative
1306	mEN356	chr1:38196744-38201861	Negative
1333	mEN357	chr1:39945533-39950689	**7/9 forebrain, 7/9 cranial nerve, 7/9 dorsal root ganglion**
1307	mEN358	chr13:34285394-34290493	**3/5 forebrain, 3/5 midbrain, 3/5 hindbrain**
1308	mEN359	chr4:97647212-97651215	Negative
1309	mEN360	chr11:117343025-117348116	**8/8 forebrain**
1310	mEN362	chr12:12707412-12712118	**5/6 forebrain, 5/6 midbrain**
1311	mEN363	chr4:62611143-62615332	Negative
1328	mEN366	chr2:101589988-101594341	**8/9 forebrain, 8/9 midbrain, 6/9 limb, 6/9 shoulder**
1312	mEN367	chr2:103623986-103627532	**3/5 forebrain, 4/5 hindbrain**
1334	mEN368	chr13:84781772-84786465	**5/10 forebrain**
1329	mEN369	chr18:34131298-34134370	**8/10 nose, 7/10 neck**
1313	mEN373	chr2:130489314-130493856	**3/6 forebrain**
1316	mEN381	chr6:93818356-93823383	**4/9 forebrain, 4/9 midbrain, 4/9 hindbrain**
1314	mEN382	chr6:91144563-91149338	**7/7 forebrain, 7/7 midbrain, 7/7 hindbrain, 4/7 trigeminal V (ganglion, cranial)**
1315	mEN383	chr16:23502808-23507356	**7/8 forebrain, 7/8 hindbrain, 4/8 neural tube**
1317	mEN388	chr1:97538497-97542741	**3/4 forebrain, 3/4 midbrain, 3/4 hindbrain, 3/4 neural tube**
1336	mEN391	chr8:87151207-87154296	**3/4 forebrain**
1338	mEN395	chr12:5266438-5269568	**8/10 ear**
1339	mEN396	chr16:37812647-37815565	Negative
1340	mEN397	chr5:77486940-77489925	**4/9 forebrain, 5/9 hindbrain, 7/9 limb**
1364	mEN400	chr6:112813562-112816924	Negative
1365	mEN401	chr3:63869819-63872427	Negative

1341	mEN402	chr14:73233956-73236326	**6/6 forebrain, 6/6 midbrain, 6/6 hindbrain, 6/6 limb, 3/6 blood vessel**
1366	mEN403	chr5:118665477-118668878	Negative
1348	mEN405	chr11:107762173-107764184	**11/11 abdomen**
1367	mEN406	chr9:95812717-95815609	**3/8 midbrain, 5/8 hindbrain, 7/8 ear**
1368	mEN409	chr2:117427080-117430606	**3/6 forebrain**
1349	mEN410	chr11:77924762-77927516	Negative
1369	mEN411	chr1:158265467-158268046	**3/4 midbrain, 3/4 hindbrain, 3/4 neck**
1370	mEN412	chr3:76465722-76469421	**7/7 forebrain, 4/7 midbrain, 4/7 hindbrain**
1371	mEN413	chr9:13697970-13700760	**6/6 Hindbrain, 3/6 neural tube**
1372	mEN414	chr1:75288287-75291172	**5/12 forebrain**
1342	mEN415	chr1:13003747-13006556	Negative
1345	mEN420	chr4:24216914-24220803	Negative
1346	mEN421	chr2:166019657-166023462	**4/5 midbrain**
1375	mEN424	chr2:168693119-168695892	**4/5 hindbrain, 3/5 neural tube**
1376	mEN425	chr13:12502078-12504879	**4/5 forebrain, 4/5 midbrain, 4/5 hindbrain, 4/5 eye, 4/5 neural tube**
1347	mEN429	chrX:99566578-99569308	**5/8 midbrain**
1389	mEN432	chr17:4038923-4041381	Negative
1406	mEN439	chr9:120601909-120604533	**5/8 midbrain**
1391	mEN440	chr2:132426454-132429102	**5/5 forebrain, 4/5 nose, 3/5 heart**
1398	mEN442	chr5:99272413-99275239	Negative
1392	mEN444	chr3:98092572-98095417	Negative
1401	mEN445	chr7:135137921-135140618	**3/4 forebrain, 3/4 midbrain**
1393	mEN448	chr12:79795794-79798372	**5/7 blood vessels**
1386	mEN451	chr6:114802640-114805326	Negative
1394	mEN453	chr8:116163758-116166268	**4/7 facial mesenchyme, 6/7 hindbrain, 7/7 neural tube**
1402	mEN454	chr2:170836158-170839441	**6/12 heart**
1403	mEN456	chr13:39876693-39879433	**6/6 forebrain, 5/6 facial mesenchyme, 6/6 neural tube, 5/6 midbrain, 6/6 hindbrain**
1390	mEN458	chr18:69546507-69549364	Negative
1395	mEN462	chrX:22897289-22900007	**10/11 forebrain, 10/11 neural tube**
1388	mEN463	chr3:51907571-51910645	**7/8 forebrain, 7/8 hindbrain, 7/8 eye,**

			7/8 midbrain, 6/8 heart, 5/8 ear, 5/8 nose, 5/8 brancial arch
1396	mEN464	chr7:6850507-6853396	Negative
1397	mEN465	chr7:76437642-76440363	**3/5 forebrain**
1399	mEN466	chrX:101985615-101988142	Negative
1387	mEN467	chr4:132032113-132036152	**4/5 forebrain**
1405	mEN468	chr7:134677970-134680502	Negative
1400	mEN469	chr15:30511450-30513962	**5/7 Trigeminal V (ganglion, cranial), 5/7 tail**
1318	mEN472	chr6:50354039-50357303	Negative
1319	mEN473	chr18:5185222-5188225	Negative
1330	mEN474	chr13:68571134-68575350	**5/7 heart, 3/7 abdomen**
1320	mEN475	chr7:80118608-80122266	Negative
1321	mEN476	chr6:39541755-39546349	**3/4 heart, 3/4 nose, 3/4 shoulder**
1352	mEN478	chr16:32852044-32856284	**9/9 heart**
1353	mEN480	chr6:145455263-145460084	Negative
1322	mEN481	chr18:65514203-65517793	**5/6 midbrain**
1323	mEN484	chr11:98901653-98906641	**5/7 abdomen**
1324	mEN485	chr2:84517965-84520803	**6/10 abdomen**
1331	mEN487	chr18:61348779-61352228	Negative
1335	mEN488	chr8:78740348-78743565	**4/8 heart, 4/8 branchial arch**
1337	mEN489	chr9:41071632-41074867	**3/6 liver**
1354	mEN492	chr2:33841463-33845838	Negative
1355	mEN495	chr1:75405116-75409810	Negative
1343	mEN499	chr11:54878925-54883929	**4/4 heart**
1344	mEN500	chr4:57536131-57540163	**5/6 heart**
1325	mEN502	chr2:31004939-31008077	Negative
1326	mEN509	chr17:30548540-30552550	**4/5 heart**
1327	mEN510	chr3:121735097-121737629	Negative
1350	mEN514	chr18:39229300-39231539	Negative
1407	mEN515	chr7:109706812-109711678	**5/6 heart, 6/6 somite**
1351	mEN518	chr19:53411035-53413469	**7/9 facial mesenchyme, 5/6 ear**
1377	mEN521	chr2:156813760-156816411	**3/5 somite**
1356	mEN524	chr5:101966725-101970386	Negative

1378	mEN526	chr9:21556521-21559582	Negative
1357	mEN527	chr1:31101599-31104444	Negative
1381	mEN528	chr19:10659775-10663888	Negative
1358	mEN530	chr1:68779329-68782031	Negative
1379	mEN531	chr7:34265554-34269796	**8/9 heart, 8/9 limb, 4/9 eye**
1380	mEN532	chr2:45053937-45057992	Negative
1382	mEN534	chr10:69643182-69647247	Negative
1359	mEN535	chr12:79968630-79971892	**4/9 heart, 8/9 branchial arch, 5/9 abdomen**
1360	mEN536	chr3:122032210-122035024	Negative
1361	mEN539	chr16:37892144-37895218	**7/9 heart, 8/9 forebrain, 9/9 limb, 5/9 blood vessels**
1384	mEN543	chr11:103049822-103053302	Negative
1383	mEN545	chr6:50336190-50338926	Negative
1362	mEN546	chr8:11356668-11359383	Negative
1363	mEN548	chr1:127754802-127759066	Negative
1385	mEN549	chr8:89992683-89995450	**3/5 heart**

Table S6 Validation results for 25 putative enhancers in four different cell lines

Region	H1-hESC	HOS	A549	TZMBL
chr1:1953310-192546069	**Positive**	**Positive**	**Positive**	**Positive**
chr2:231809337-231809988	Negative	**Positive**	**Positive**	**Positive**
chr9:134224987-134225644	-	-	-	-
chr11:65679112-61679919	**Positive**	**Positive**	**Positive**	**Positive**
chr12:125039037-125040700	**Positive**	**Positive**	**Positive**	**Positive**
chr13:113921562-113922944	**Positive**	**Positive**	**Positive**	**Positive**
chr14:77422602-77423265	**Positive**	**Positive**	**Positive**	**Positive**
chr17:2929462-2930394	**Positive**	**Positive**	**Positive**	**Positive**
chr17:72390462-72391344	-	-	-	-
chr22:31662162-31663116	Negative	**Positive**	**Positive**	**Positive**
chr1:54839458-54841157	Negative	**Positive**	Negative	**Positive**
chr3:128150669-128152511	**Positive**	Negative	Negative	Negative
chr4:6246837-6247511	**Positive**	**Positive**	**Positive**	**Positive**
chr7:1956626-1958036	**Positive**	Negative	**Positive**	**Positive**
chr7:73448387-73448811	Negative	Negative	**Positive**	Negative
chr9:132976212-132977003	Negative	**Positive**	**Positive**	**Positive**
chr9:138892812-1338893419	**Positive**	Negative	Negative	Negative
chr11:44307337-44308437	Negative	Negative	**Positive**	Negative
chr12:52536500-52539000	Negative	Negative	Negative	Negative
chr13:24121112-24121886	**Positive**	**Positive**	**Positive**	**Positive**
chr14:75905362-75907344	**Positive**	Negative	**Positive**	Negative
chr18:12271615-12272169	Negative	**Positive**	**Positive**	**Positive**
chr19:6235287-6237180	**Positive**	Negative	**Positive**	Negative
chr22:44243837-44244786	Negative	Negative	Negative	Negative
chr22:45986287-45987069	Negative	Negative	Negative	Negative
Overall	**13/23**	**13/23**	**16/23**	**13/23**

Table S7 The fold change of gene expression as compared to control sequences in the forward as well as reverse directions for the 25 putative enhancers.

Element	H1-hESC	HOS	A549	TZMBL
chr1:1953310-192546069	3.06, 7.55	18.67, 60.75	3, 19.9	5.67, 9.67
chr2:231809337-231809988	0. 1.06	6.33, 3.83	3.21, 0.48	3.58, 2.08
chr9:134224987-134225644	-	-	-	-
chr11:65679112-61679919	2.86, 2.45	8.17,25.83	14.2, 2.42	5.17, 9.75
chr12:125039037-125040700	0, 2.24	11.17, 11.67	1.31, 4.9	6.58, 8.25
chr13:113921562-113922944	1.20, 4.49	18.67, 9.83	6.1, 1.1	8.25, 5.75
chr14:77422602-77423265	11.84, 2.04	34.58, 3.5	0.24, 0.24	10, 0.55
chr17:2929462-2930394	0, 11.63	0.92, 37.5	0.71, 54.5	0.33, 6.92
chr17:72390462-72391344	-	-	-	-
chr22:31662162-31663116	0, 1.02	1.83, 7.0	2.4, 2.1	0.92, 1.25
chr1:54839458-54841157	0, 1.80	10.58, 1.33	1.8, 0.12	2.58, 0.12
chr3:128150669-128152511	2.24, 1.78	2.17, 1.42	0.24, 0.25	0.48, 1.17
chr4:6246837-6247511	11.63, 0.88	40.75, 1	43.75, 0.79	5.5, 0.16
chr7:1956626-1958036	6.53, 0	0.83, 1.19	29.73, 1.11	14.3, 0
chr7:73448387-73448811	0, 1.73	0.97, 1.36	1.64, 2.19	0.57, 1.21
chr9:132976212-132977003	0.90, 0.88	0.51, 6.71	0.36, 14.93	0.93, 6.3
chr9:138892812-1338893419	1.82, 0	0.66, 0.51	0.88, 0.72	0.46, 0.34
chr11:44307337-44308437	0, 0	0.89, 0.85	0, 5.48	0, 1.2
chr12:52536500-52539000	0. 0.42	0.16, 1.34	0.53, 0.52	1, 0.93
chr13:24121112-24121886	3.24, 0.39	4.79, 7.34	11.09, 38.36	4.8, 4.6
chr14:75905362-75907344	4.06, 0	2.05, 1.78	7.34, 2.19	1, 1.1
chr18:12271615-12272169	0.42, 0.44	2.74, 3.15	6.44, 4.38	2.5, 4.1
chr19:6235287-6237180	6.72, 0.97	1.15, 0.16	23.97, 0.68	0.81, 0
chr22:44243837-44244786	0.82, 0.89	0.12, 0	0.20, 0.01	0.99, 1.02
chr22:45986287-45987069	1.88, 0.46	0.19, 0	0.16, 0.07	1.08, 0.87

Figures and Captions:

Figure S1

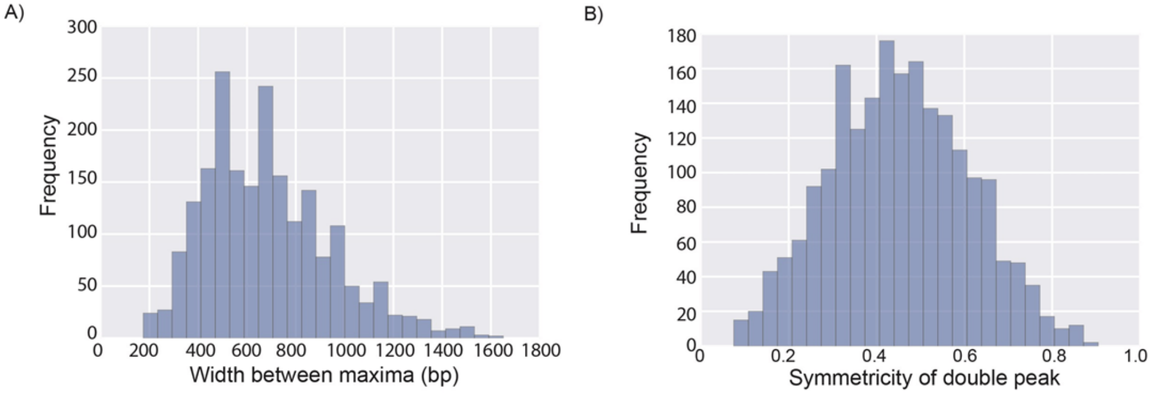


Figure S1: Variability in double peak pattern. A) The frequency of distance between the two maxima in a double peak flanking active STARR-seq peaks is plotted. B) The symmetricity of the double peak pattern is plotted. The ratio of the distance between the two peaks to the ratio between one of the maxima and the minima is plotted. While there is large amount of variability in the distance between the two peaks (mostly between 300-1100 bp), the trough in the double peak tends to occur in the center of the two peaks.

Figure S2

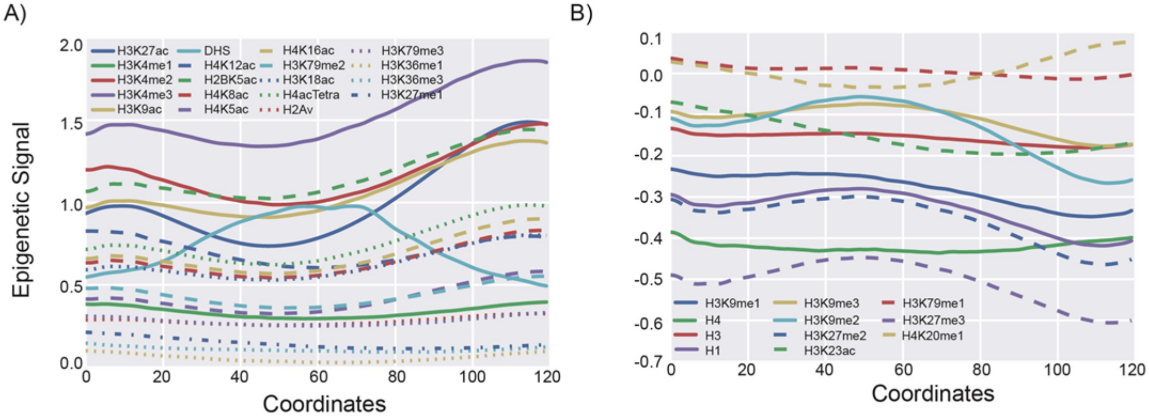


Figure S2: Metaprofile for different epigenetic marks. The metaprofile around active STARR-seq peaks is plotted for different epigenetic marks. Histone marks that are enriched near STARR-seq peaks display the characteristic double peak pattern shown in A) due to the depletion of histone proteins at active regulatory regions. In addition, DHS displays a single peak at the center of these regulatory regions as shown in A). B) On the other hand, no such double peak pattern is observed on depleted histone marks at STARR-seq peaks.

Figure S3

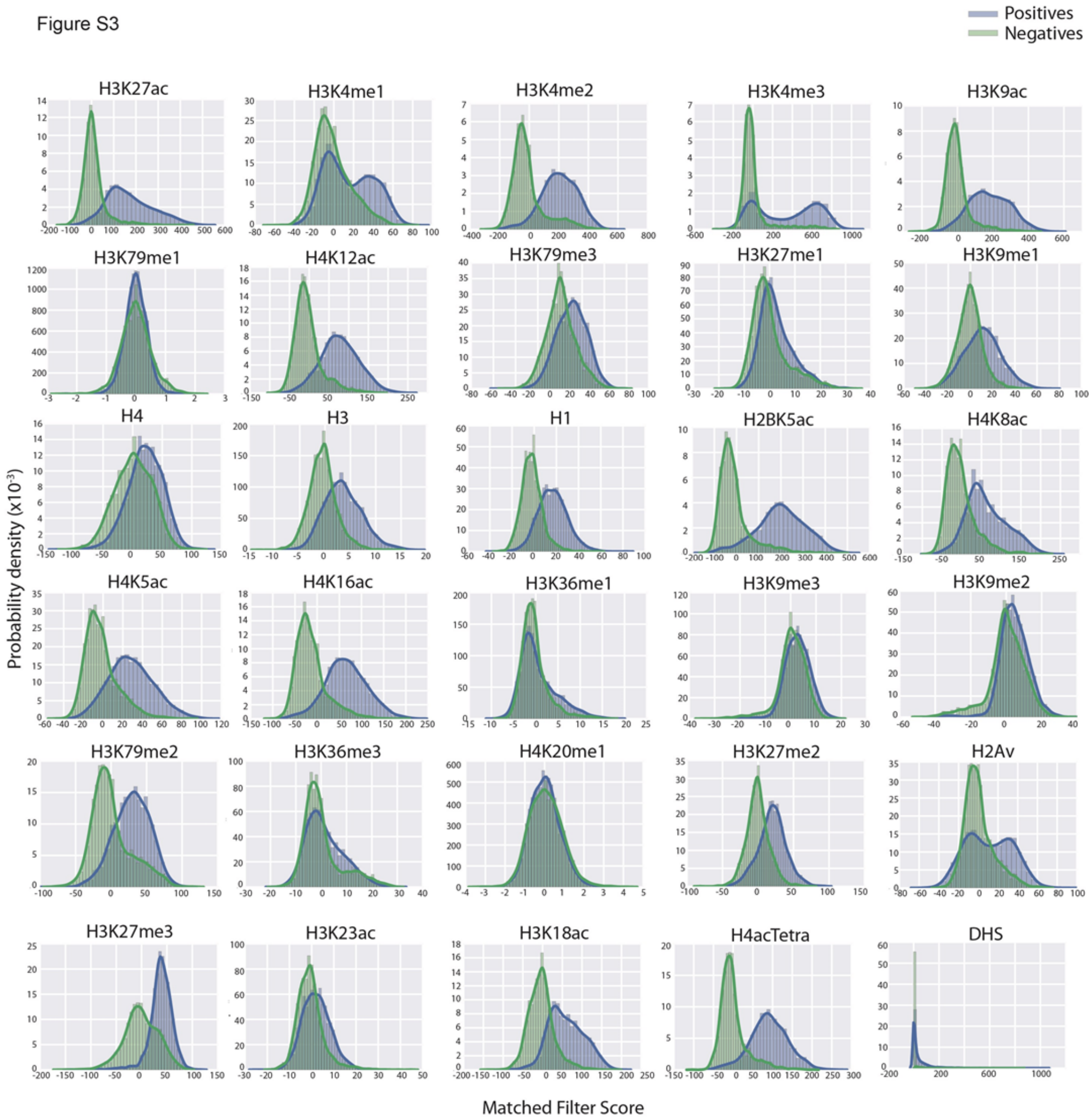


Figure S3: Histogram of matched filter scores. The probability density of matched filter scores for different epigenetic marks for STARR-seq peaks (positives) and random regions of the genome (negatives) with H3K27ac signal. In most cases, the matched filter scores for positives and negatives are Gaussian curves. The amount of overlap between these two curves determines the accuracy of the matched filter for predicting STARR-seq peaks using the matched filters for the corresponding epigenetic feature.

Figure S4

A)

Feature	AUROC	AUPR
H3K27ac	0.92 (0.83)	0.72 (0.63)
H3K9ac	0.89 (0.77)	0.52 (0.39)
DHS	0.86 (0.77)	0.58 (0.67)
H3K4me2	0.87 (0.75)	0.41 (0.34)
H3K4me3	0.73 (0.64)	0.32 (0.28)
H3K4me1	0.80 (0.72)	0.46 (0.39)

B)

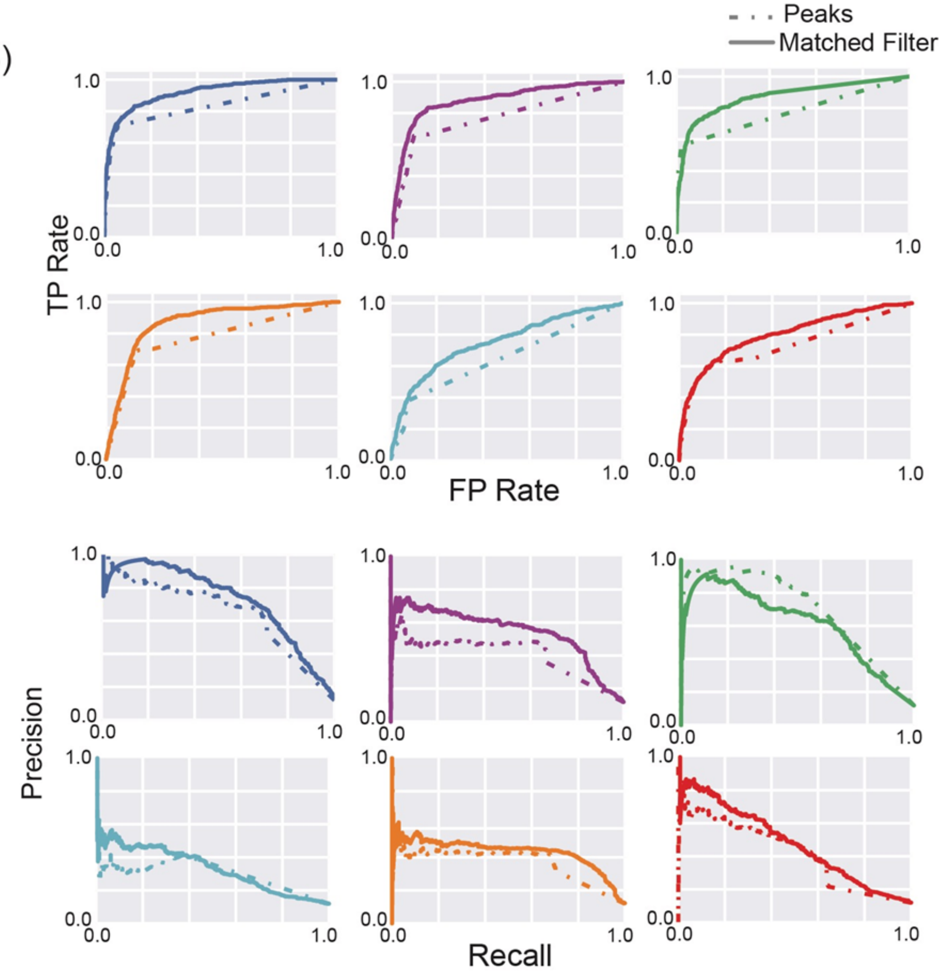


Figure S4: Accuracy of matched filter and peak-based models. The performance of the matched filters of different epigenetic marks and the peak-based models for predicting all STARR-seq peaks is compared here using 10-fold cross validation. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (multiple core promoters) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for the matched filter model. B) The individual ROC and PR curves for each matched filter and the peak-based model are shown.

Figure S5

A)

Model	AUROC	AUPR
Random Forest	0.96 (0.95)	0.91 (0.79)
Ridge Regression	0.95 (0.94)	0.90 (0.77)
Linear SVM	0.96 (0.95)	0.91 (0.78)
Naive Bayes	0.95 (0.93)	0.89 (0.72)

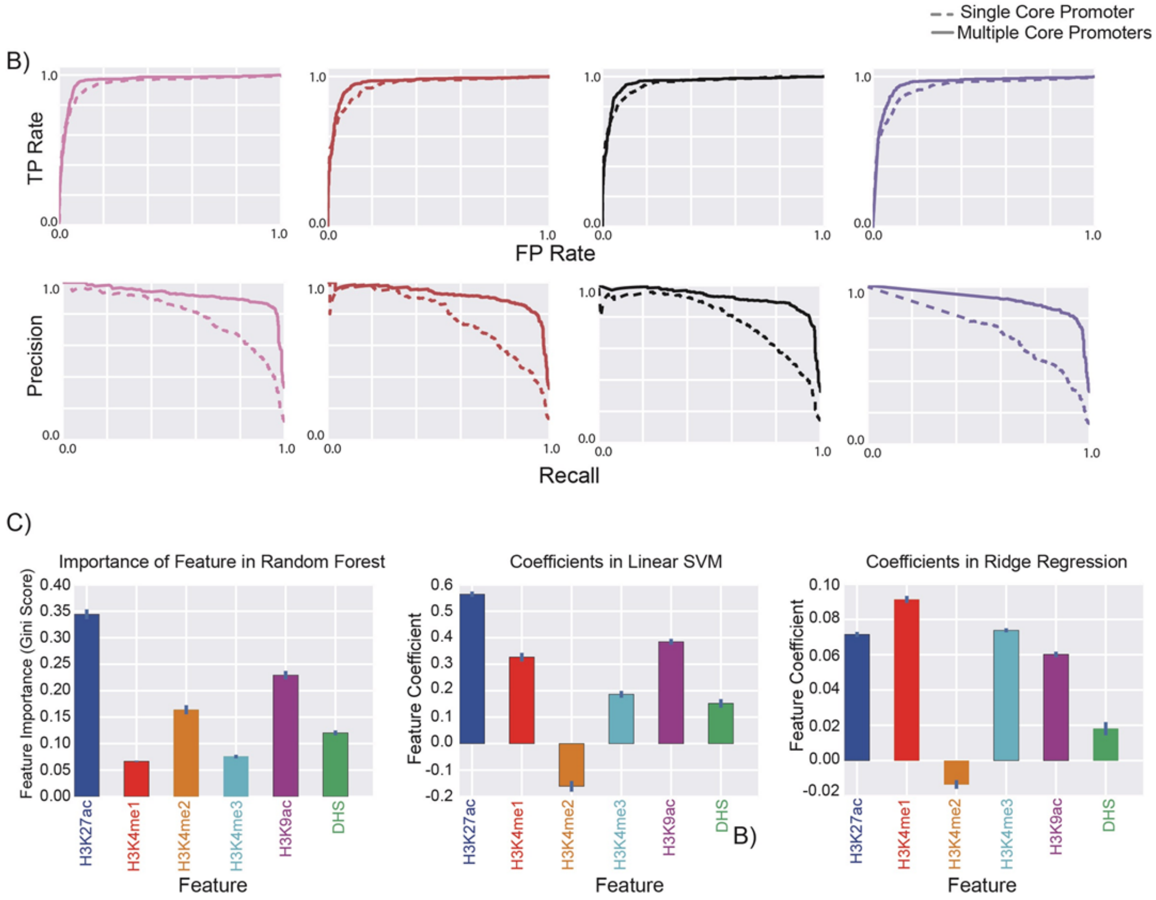


Figure S5: Comparison of different statistical models. The performance of the different statistical models to integrate the information from six epigenetic features is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (single core promoter) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting STARR-seq peaks identified after combining multiple core promoters. B) The individual ROC and PR curves for each statistical model. C) The contribution of the matched filter score for each epigenetic feature to the different integrated models.

Figure S6

A)

Feature	AUROC	AUPR
H3K27ac	0.88 (0.94)	0.78 (0.87)
H3K9ac	0.86 (0.94)	0.56 (0.86)
H3K4me2	0.84 (0.92)	0.53 (0.79)
H3K4me3	0.58 (0.91)	0.28 (0.84)
H3K4me1	0.89 (0.58)	0.74 (0.44)
Random Forest	0.91 (0.94)	0.81 (0.90)
Ridge Regression	0.93 (0.96)	0.84 (0.90)
Linear SVM	0.92 (0.95)	0.84 (0.90)
Naive Bayes	0.92 (0.96)	0.82 (0.91)

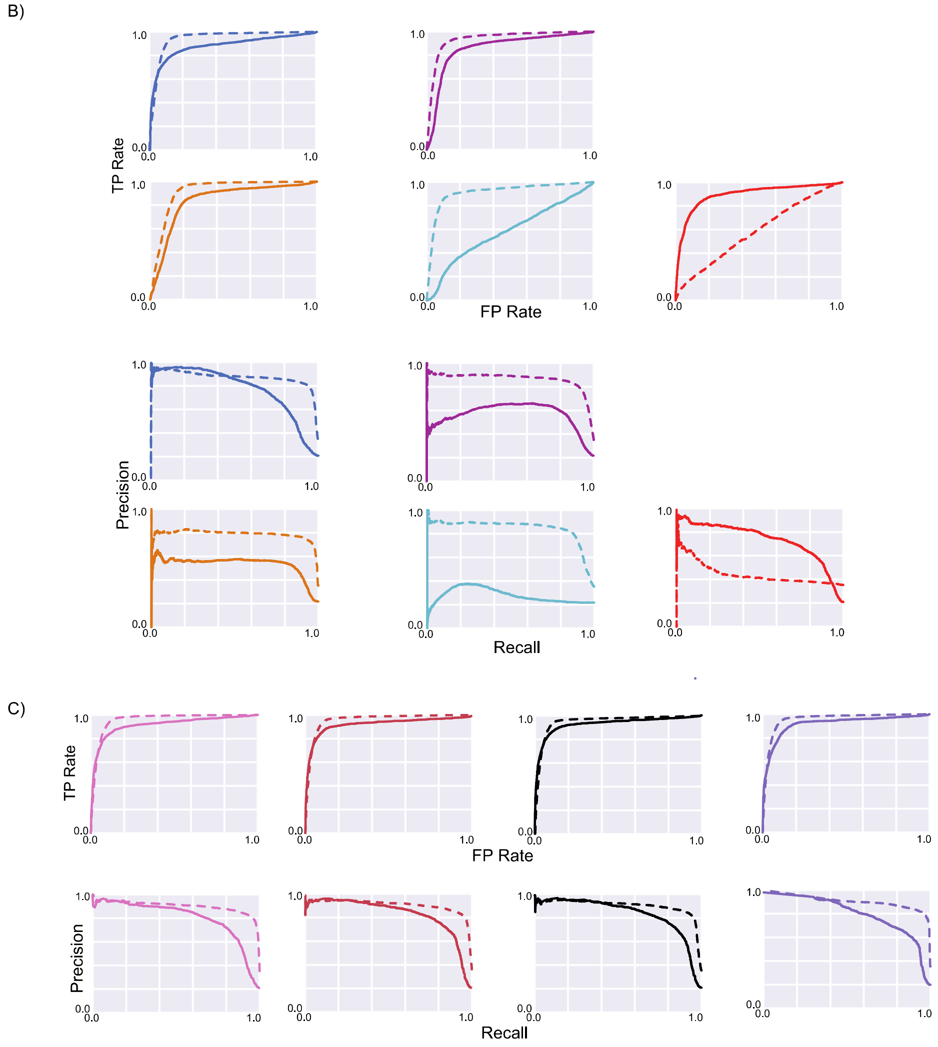


Figure S6: Transferability of models across cell-lines. The performance of the BG3-trained matched filters of different epigenetic marks and statistical models for predicting active promoters and enhancers are compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

Figure S7

A)

Model	AUROC	AUPR
Random Forest	0.97 (0.98)	0.92 (0.84)
Ridge Regression	0.96 (0.97)	0.92 (0.81)
Linear SVM	0.97 (0.97)	0.93 (0.83)
Naive Bayes	0.94 (0.95)	0.90 (0.72)

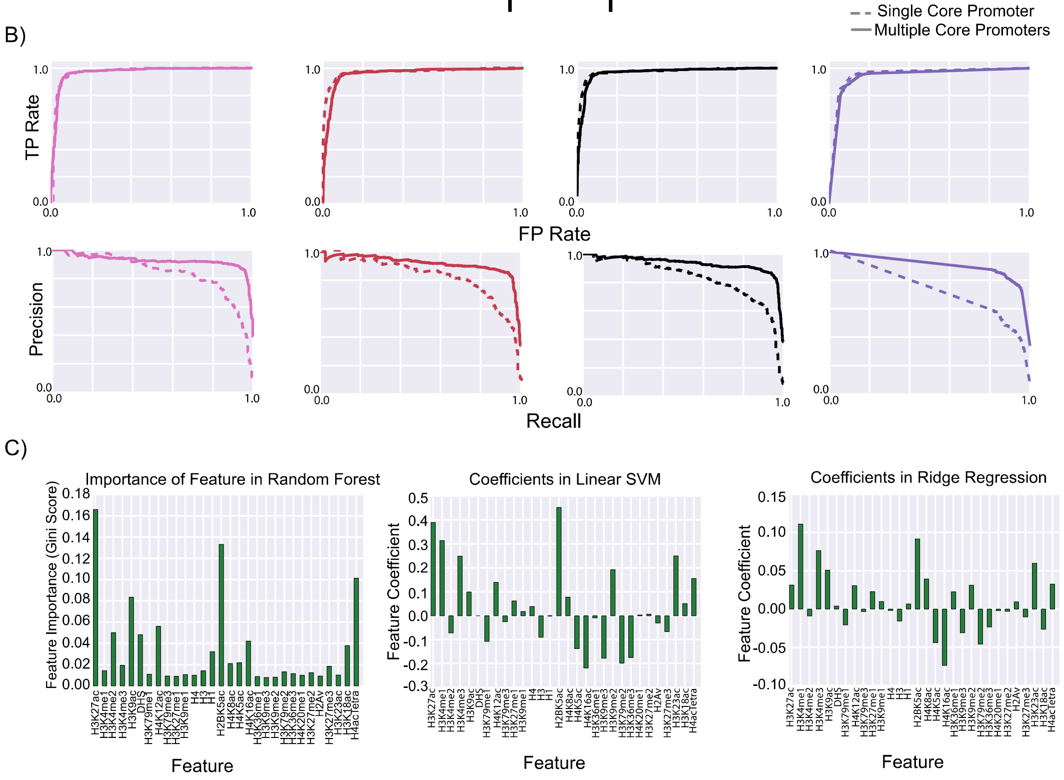


Figure S7: Comparison of different statistical models for 30-feature model. The performance of the different statistical models to integrate the information from 30 epigenetic features is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the STARR-seq peaks (single core promoter) with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting STARR-seq peaks identified after combining multiple core promoters. B) The individual ROC and PR curves for each statistical model. C) The contribution of the matched filter score for each epigenetic feature to the different integrated models.

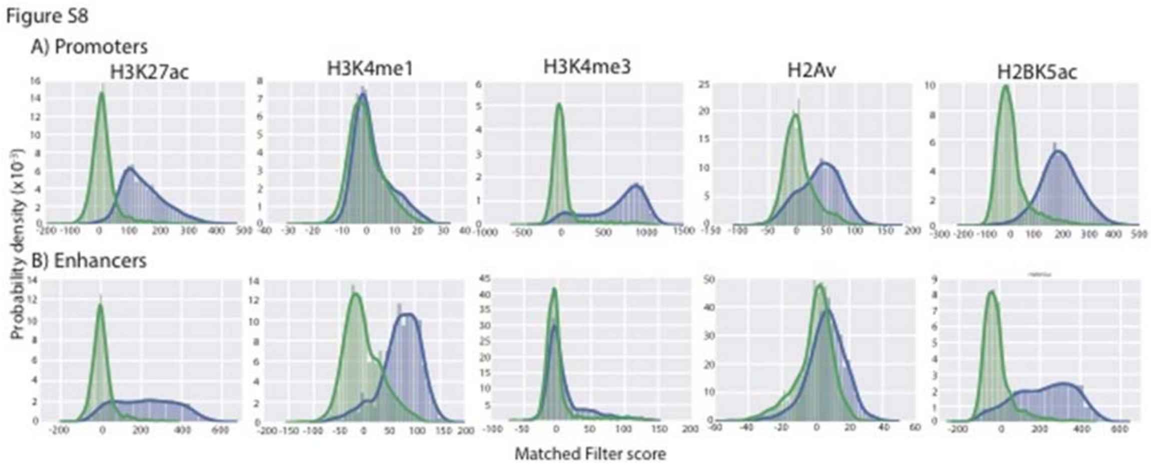


Figure S8: Histogram of matched filter scores for chosen features in promoters and enhancers. A) The histogram of matched filter scores for small set of epigenetic features on promoters is compared to random regions of the genome. B) The histogram of matched filter scores for small set of epigenetic features on enhancers is compared to random regions of the genome.

Figure S9

A)

Model	AUROC	AUPR
Random Forest	0.95 (0.97)	0.66 (0.78)
Ridge Regression	0.94 (0.97)	0.65 (0.78)
Linear SVM	0.94 (0.97)	0.66 (0.78)
Naive Bayes	0.92 (0.97)	0.54 (0.79)

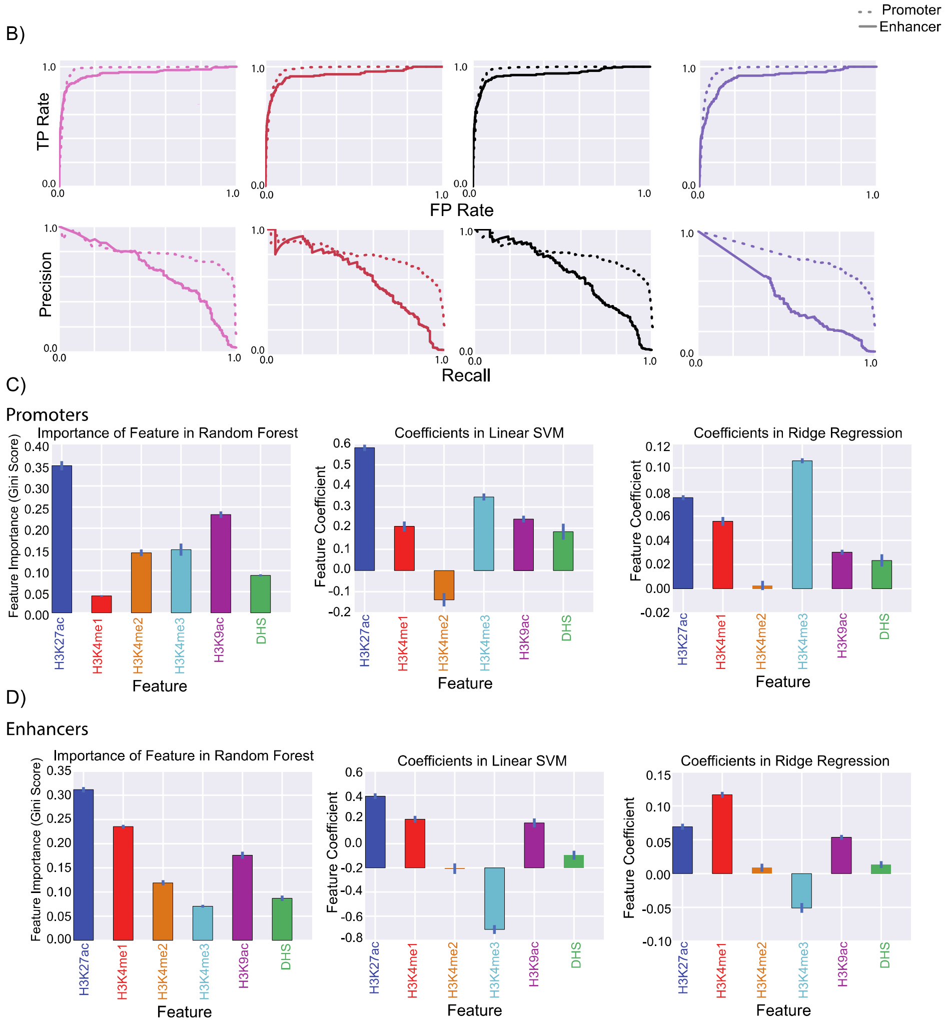


Figure S9: Comparison of different statistical models for predicting enhancers and promoters. The performance of the different statistical models to integrate the information from six epigenetic features for promoter and enhancer prediction is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the promoters with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting enhancers. The promoters and enhancers from multiple STARR-seq experiments with different core promoters are merged in this analysis. B) The individual ROC and PR curves for each statistical model is shown. The contribution of the matched filter score for each epigenetic feature to the different integrated models for promoter prediction (C) and enhancer prediction (D) are shown.

Figure S10

A)

Feature	AUROC	AUPR
H3K27ac	0.94	0.92
H3K9ac	0.93	0.92
DHS	0.89	0.89
H3K4me2	0.91	0.87
H3K4me3	0.91	0.90
H3K4me1	0.57	0.59
Random Forest	0.85	0.84
Ridge Regression	0.82	0.80
Linear SVM	0.79	0.80
Naive Bayes	0.95	0.93

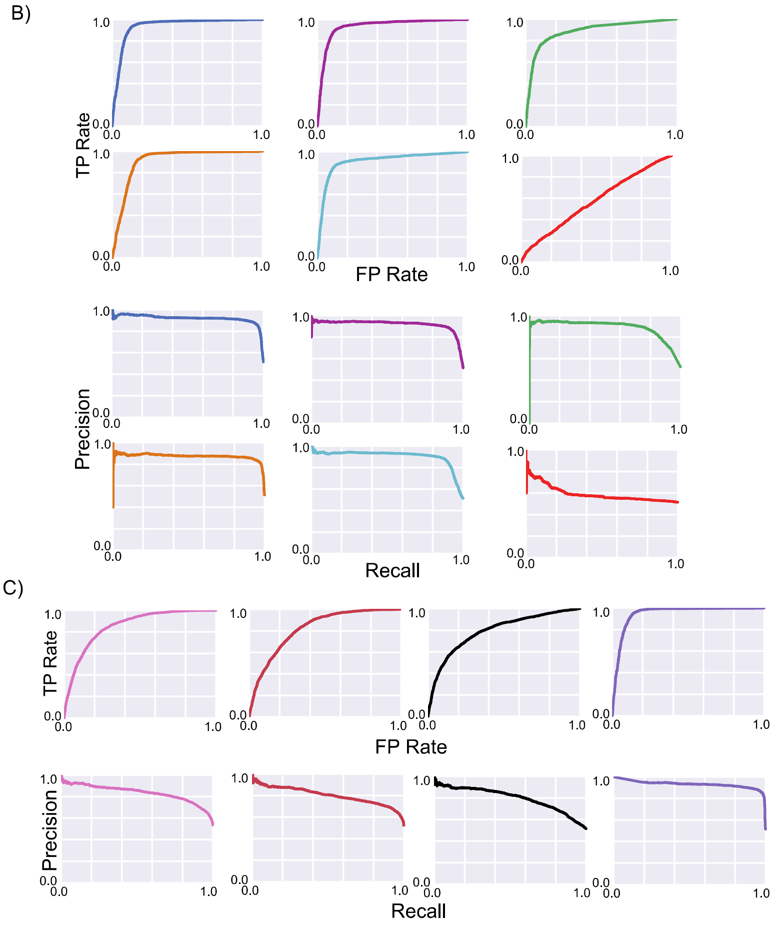


Figure S10: Accuracy of enhancer-trained matched filter and statistical models for promoter prediction. The performance of the enhancer-trained matched filters of different epigenetic marks and statistical models for predicting active promoters is compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

Figure S11

A)

Feature	AUROC	AUPR
H3K27ac	0.88	0.86
H3K9ac	0.86	0.73
DHS	0.82	0.77
H3K4me2	0.83	0.70
H3K4me3	0.58	0.46
H3K4me1	0.89	0.83
Random Forest	0.91	0.82
Ridge Regression	0.89	0.80
Linear SVM	0.90	0.86
Naive Bayes	0.88	0.83

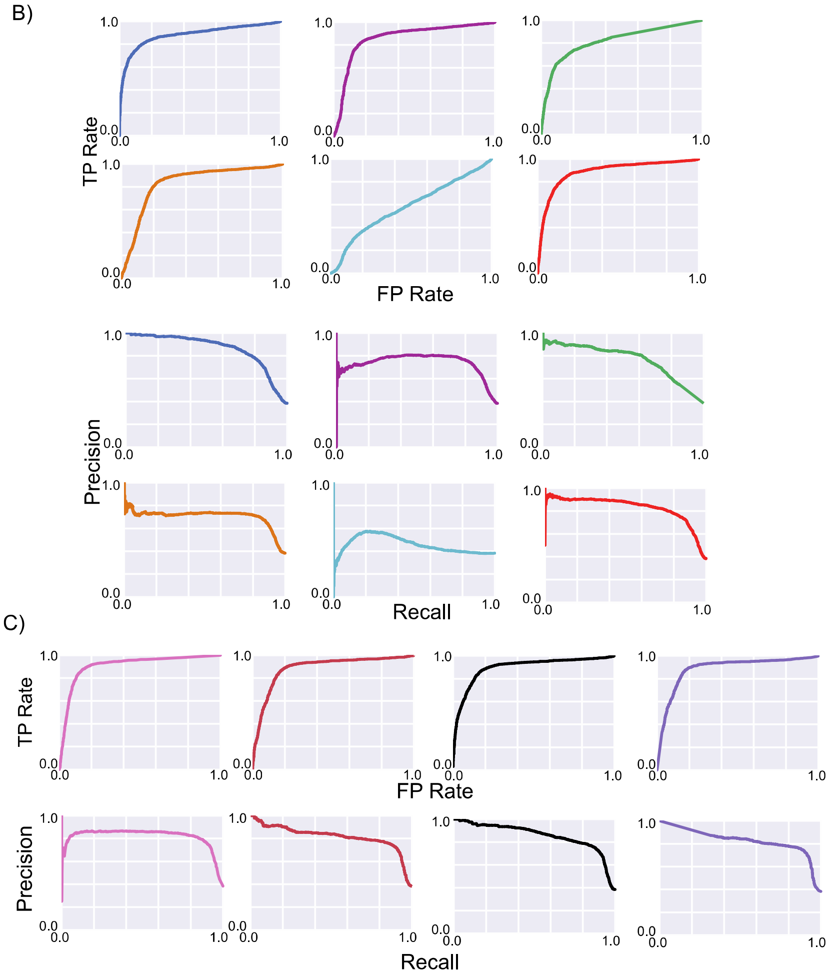


Figure S11: Accuracy of promoter-trained matched filter and statistical models for enhancer prediction. The performance of the promoter-trained matched filters of different epigenetic marks and statistical models for predicting active enhancers is compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

Figure S12

A)

Model	AUROC	AUPR
Random Forest	0.95 (0.98)	0.73 (0.75)
Ridge Regression	0.94 (0.97)	0.69 (0.81)
Linear SVM	0.95 (0.98)	0.74 (0.81)
Naive Bayes	0.91 (0.95)	0.62 (0.77)

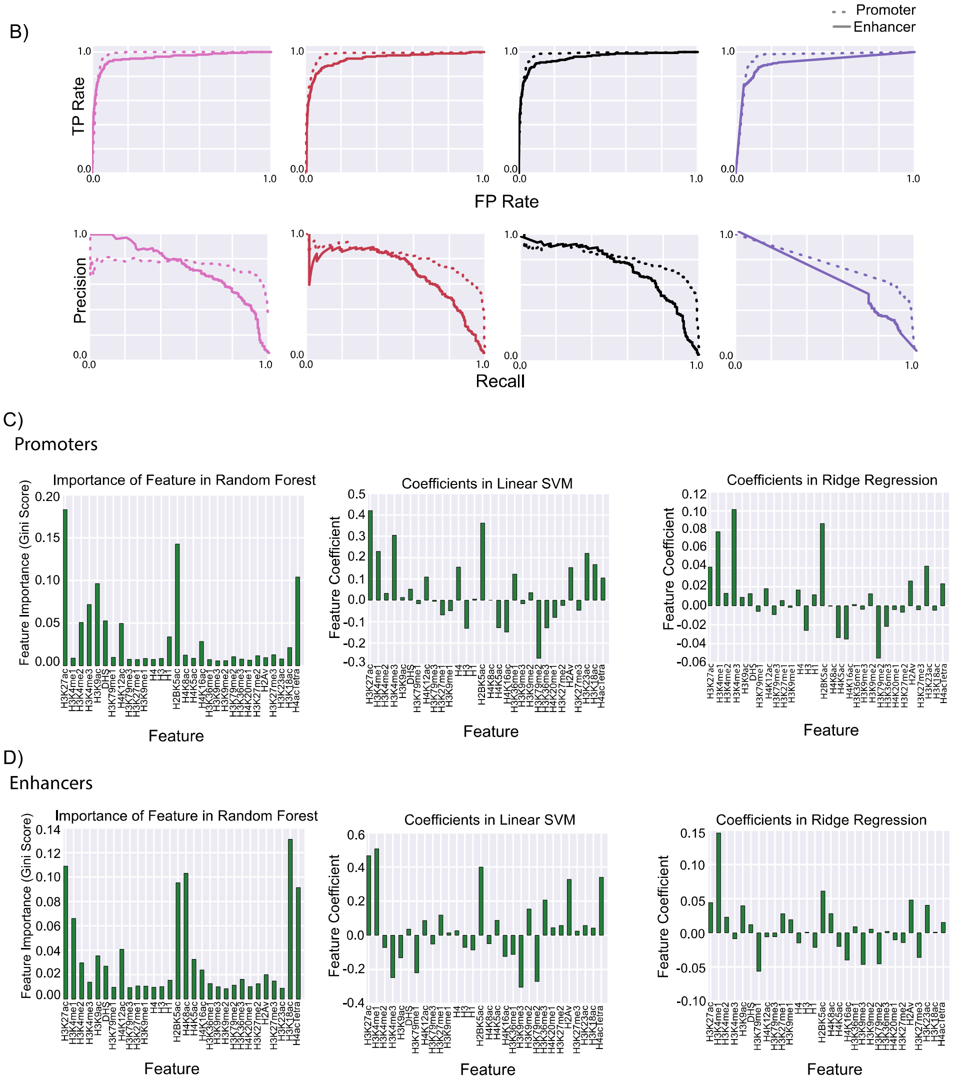


Figure S12: Comparison of different statistical models for predicting enhancers and promoters. The performance of the different statistical models to integrate the information from thirty epigenetic features for promoter and enhancer prediction is shown. A) The numbers within the parentheses refer to the AUROC and AUPR for predicting the promoters with histone peaks while the numbers outside the parentheses refer to the AUROC and AUPR for predicting enhancers. The promoters and enhancers from multiple STARR-seq experiments with different core promoters are merged in this analysis. B) The individual ROC and PR curves for each statistical model is shown. The contribution of the matched filter score for each epigenetic feature to the different integrated models for promoter prediction (C) and enhancer prediction (D) are shown.

Figure S13

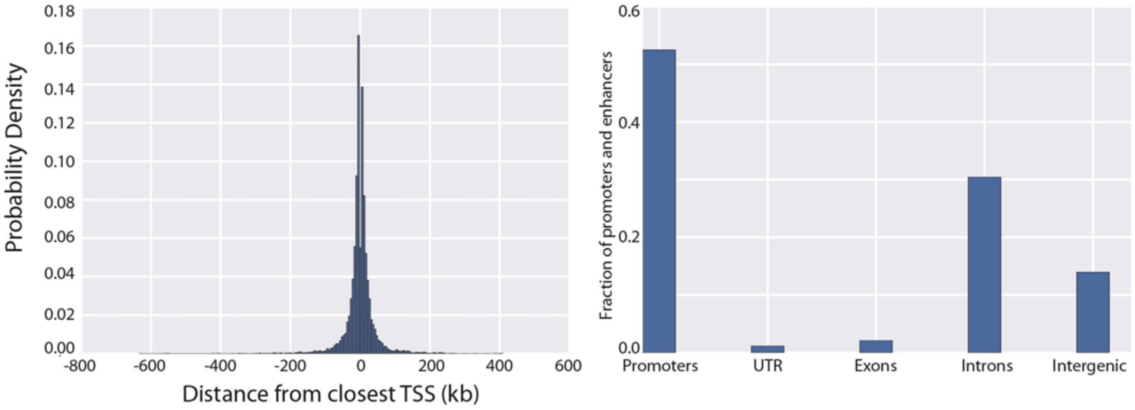


Figure S13: Location of H1-hESC predictions. A) The probability density of the distance of the predicted promoter and enhancer from the closest TSS is shown. B) The location of the enhancers and promoters on genomic elements are shown. Promoters are defined as TSS +/- 2kb. All TSS, UTR, exons, introns, and intergenic elements are calculated based on GENCODE 19 definitions [13]. A regulatory region is considered to overlap with the elements if more than 50% of the matched filter region overlaps with the corresponding element in B.

Figure S14

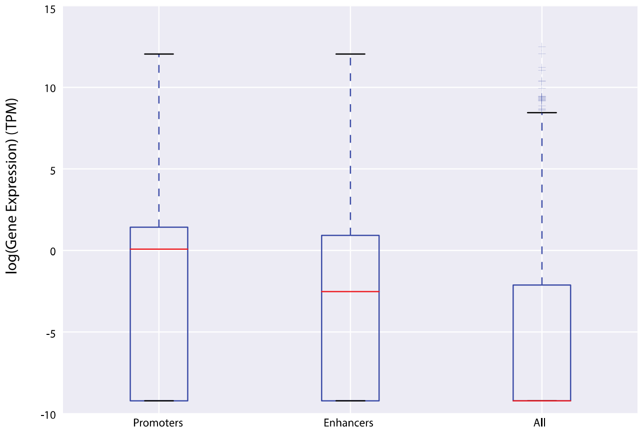


Figure S14: Gene expression of closest gene. The distribution of gene expression of gene closest to the enhancer/promoters are plotted and compared to the gene expression of all genes in H1-hESC. A Wilcoxon test shows that P-value for differences in gene expression of genes close to enhancers and promoters are significantly higher than expression of all genes in H1-hESC (< 10^{-100} each).

Figure S15

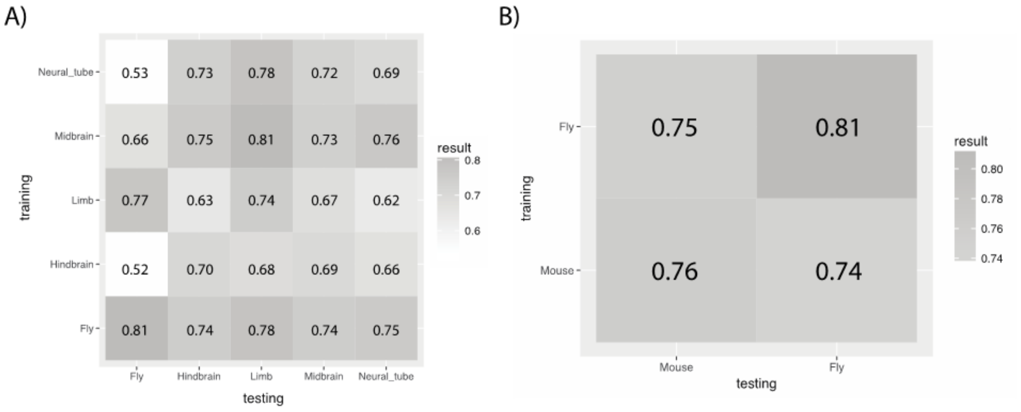


Figure S15: Cross-comparison of integrated models for enhancer prediction. To compare the performance of the integrated model trained on datasets of different sizes from different organisms, we performed cross test where the integrated model is first trained with fly STARR-seq data, cross-validated and tested on transgenic mouse assay regions. Then the model is trained in the same way with transgenic mouse assay regions, cross-validated and tested on fly S2 STARR-seq data. A) Models are trained in a cell line and tissue specific fashion. The AUROC values of each pairwise cross-validation or test are compared in the matrix. The model trained with fly STARR-seq data exhibits better performance in general. B) Assumed identical distribution of matched filter scores for active enhancer regions in each tissue in mouse, we combined the normalized matched filter scores to get a larger training set for the model. The resulting matrix demonstrated that the STARR-seq model still exhibits better performance in general.

Figure S16

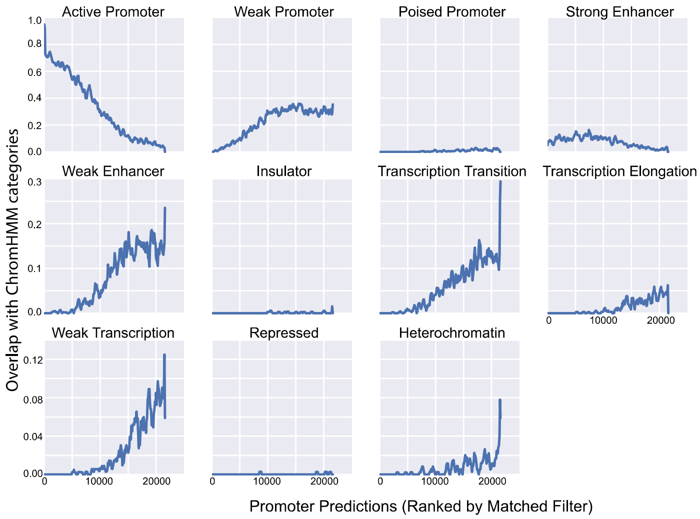


Figure S16: Overlap of predicted promoters with chromatin states predicted by ChromHMM. The promoters predicted to be active by matched filter in H1-hESC cell line are compared with the chromatin states predicted using chromHMM. Most of the matched filter promoters are also predicted to be either strong or weak promoters by chromHMM while some of the other matched filter promoters are labeled as weak enhancers or transcription related elements in chromHMM. However, very few inactive regions and insulators are predicted to be promoters by matched filter. However, the boundaries of the elements can be very different as chromHMM promoters can also be tens of kilobases in length.

Figure S17

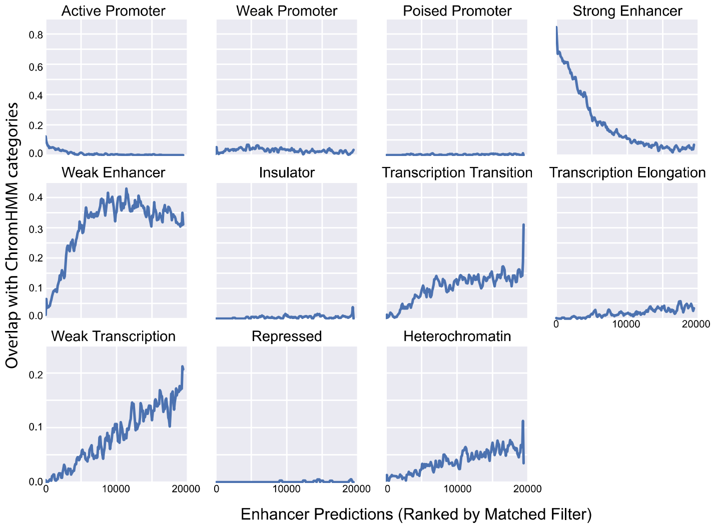


Figure S17: Overlap of predicted enhancers with chromatin states predicted by ChromHMM. The enhancers predicted to be active by matched filter in H1-hESC cell line are compared with the chromatin states predicted using chromHMM. Most of the matched filter enhancers are also predicted to be either strong or weak enhancers by chromHMM while some of the other matched filter promoters are labeled as transcription related elements in chromHMM. However, very few inactive regions and insulators are predicted to be promoters by matched filter.

Figure S18

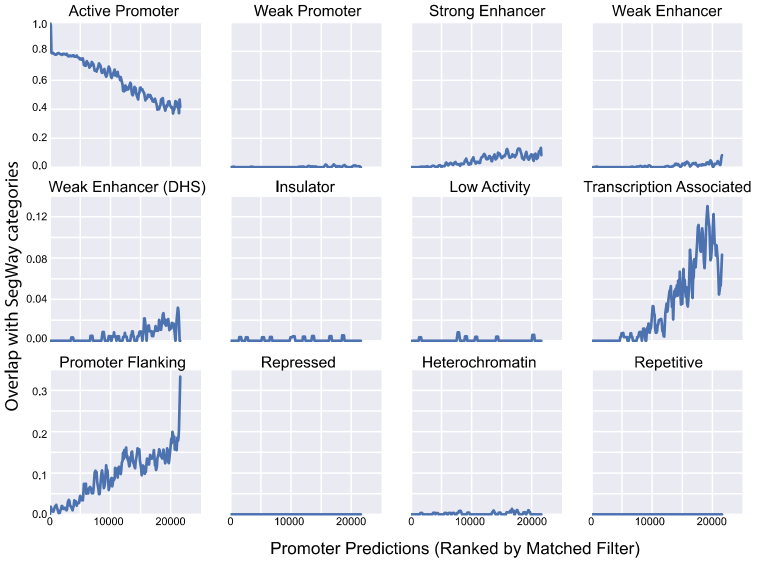


Figure S18: Overlap of predicted promoters with chromatin states predicted by SegWay. The promoters predicted to be active by matched filter in H1-hESC cell line are compared with the chromatin states predicted using SegWay. Most of the matched filter promoters are also predicted to be either active promoters by SegWay while some of the other matched filter promoters are labeled as promoter flanking or transcription related elements in SegWay. However, very few inactive regions and insulators are predicted to be promoters by matched filter. However, the boundaries of the elements can be very different.

Figure S19

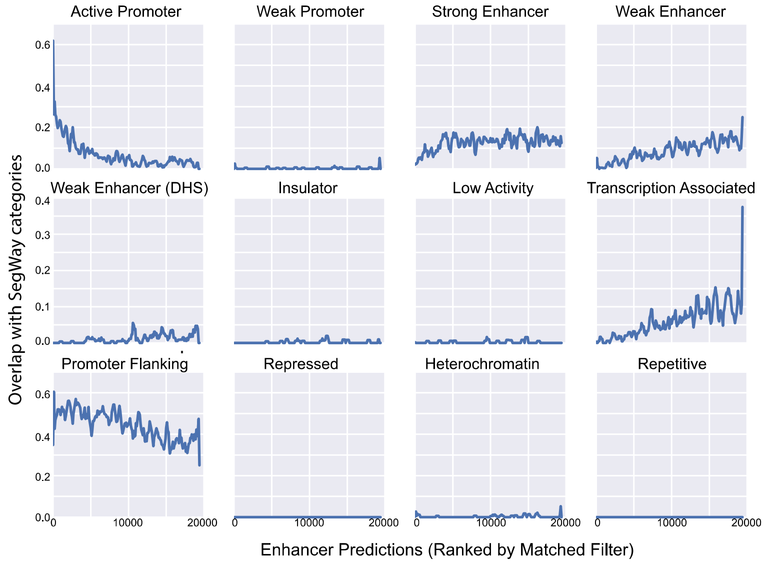


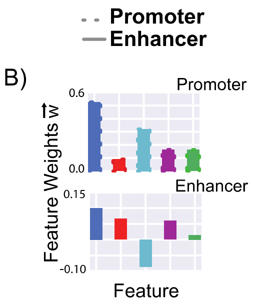
Figure S19: Overlap of predicted enhancers with chromatin states predicted by SegWay. The enhancers predicted to be active by matched filter in H1-hESC cell line are compared with the chromatin states predicted using SegWay. Most of the matched filter enhancers are also predicted to be promoters or enhancers by SegWay while some of the other matched filter enhancers are labeled as either promoter flanking or transcription related elements in SegWay. However, very few inactive regions and insulators are predicted to be promoters by matched filter.

Figure S20

Fly-based models on mouse

A)

Feature	AUROC	AUPR
H3K27ac	0.86 (0.95)	0.38 (0.71)
H3K9ac	0.80 (0.97)	0.23 (0.83)
DHS	0.90 (0.96)	0.34 (0.70)
H3K4me3	0.74 (0.97)	0.21 (0.82)
H3K4me1	0.83 (0.66)	0.27 (0.17)
Integrated	0.87 (0.98)	0.40 (0.83)



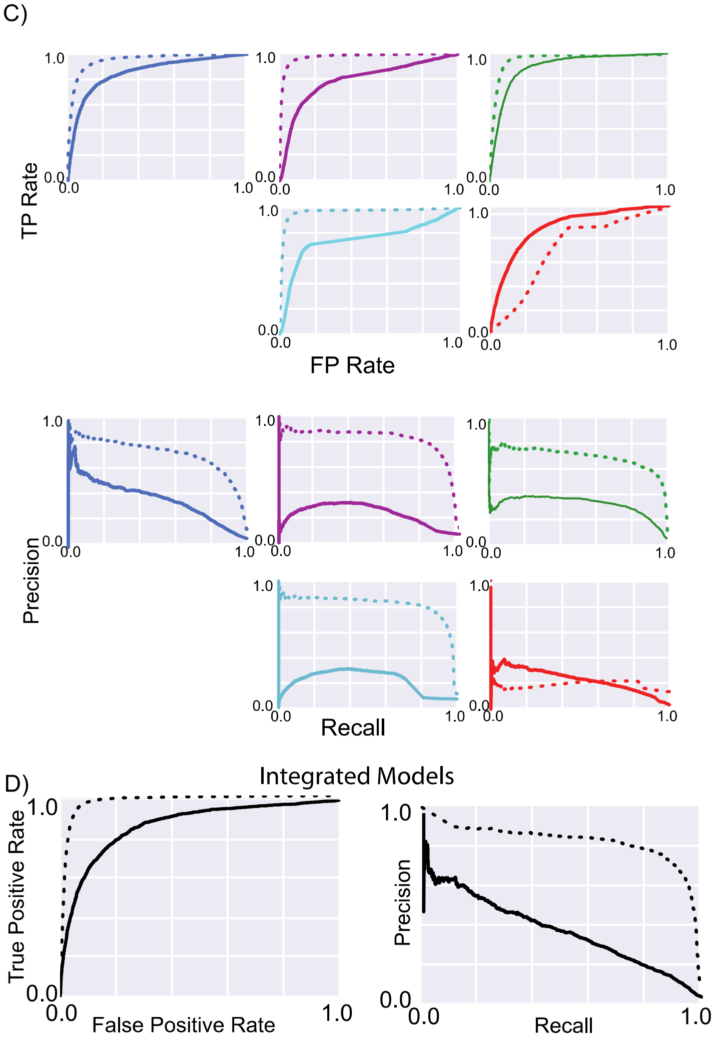


Figure S20: Accuracy of STARR-seq trained matched filter model for enhancer prediction in mouse. The performance of the fly-based matched filters and the integrated model for predicting active promoters and enhancers in mouse embryonic stem cells identified using FIREWACh. A) Similar to Figure 3, the numbers within parentheses refer to the AUROC and AUPR for predicting promoters while the numbers outside parentheses refer the performance of the models for predicting enhancers. B) The weights of the different features in the integrated models for promoter and enhancer prediction are shown. C) The individual ROC and PR curves for each matched filter and the integrated model are shown. The performance of these features and the integrated model for predicting the active promoters and enhancers identified using FIREWACh are shown.

Figure S21

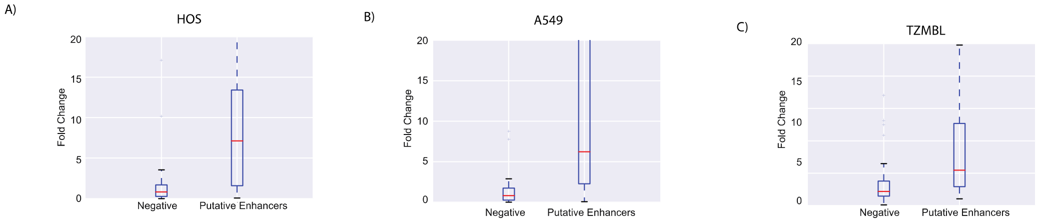


Figure S21: Activity of putative enhancers in three different cell-lines. While the enhancers were predicted in H1-hESC, the activity of these enhancers is compared in three other cell-lines and the enhancers are active in these cell-lines too.

Figure S22

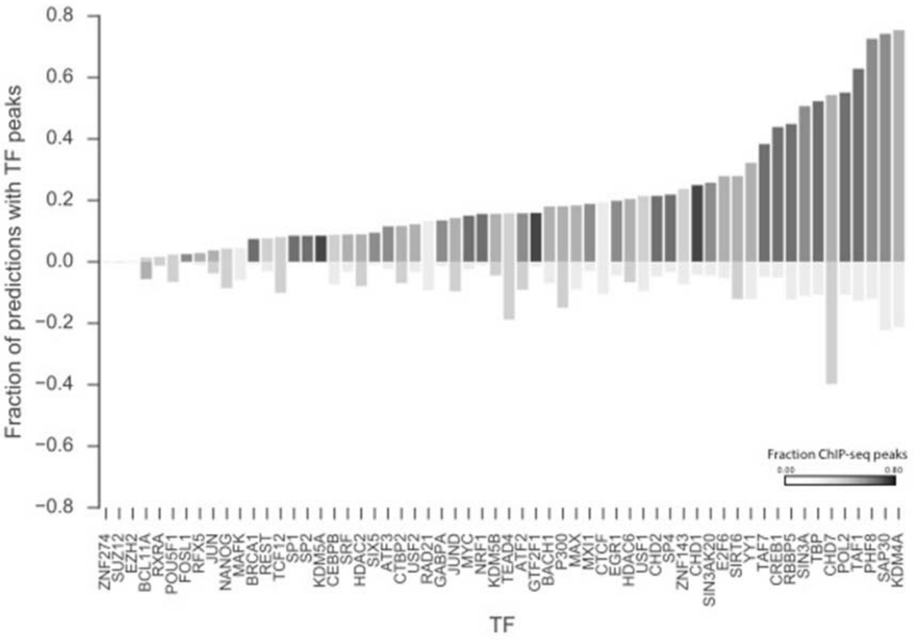


Figure S22: Overlap of TF binding site with predicted promoters/enhancers. The fraction of promoters and enhancers that overlap with different TF ChIP-seq peaks in H1-hESC are plotted. The color of the bar is plotted based on the fraction of ChIP-seq peaks for corresponding TF that overlap with the promoter/enhancer. The difference in patterns of TF binding was used to create models that distinguish enhancers from promoters (Figure 5B).

Figure S23

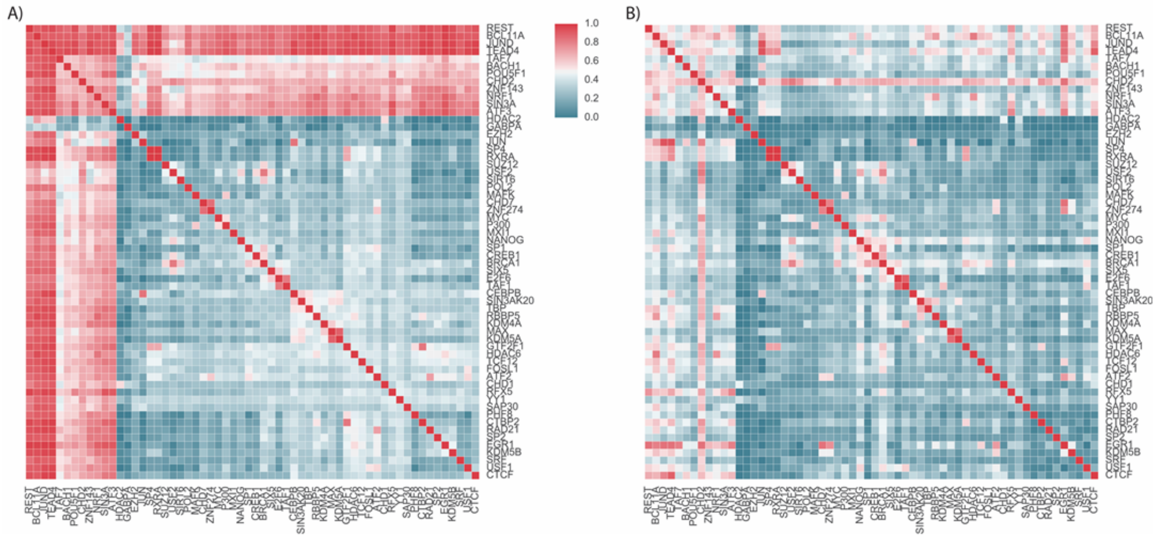


Figure S23: Patterns of co-TF binding on enhancers and promoters. The patterns of TF co-occurrence on a single matched filter prediction around promoters (A) and enhancers (B) are plotted. The differences between co-TF binding at enhancers and promoters can be used to gain some mechanistic insight into TF cooperativity.

References:

1. mod, E.C., et al., *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-97.
2. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
3. Kumar, V.B.V.K., A. Mahalanobis, and R.D. Juday, *Correlation Pattern Recognition*. 2005.
4. Blanchard, G., O. Bousquet, and P. Massaer, *Statistical performance of support vector machines.* Ann. Statist., 2008. **36**: p. 489-531.
5. Hoerl, A.E. and R.W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55--67.
6. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5--32.
7. Stuart, R. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2nd ed. 2003.
8. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825--2830.
9. Davis, J. and M. Goadrich, *The Relationship Between Precision-Recall and ROC Curves.* Proceedings of the 23rd international conference on Machine Learning, 2006: p. 233-240.
10. Kothary, R., et al., *Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice.* Development, 1989. **105**(4): p. 707-714.
11. Pennacchio, L.A., et al., *In vivo enhancer analysis of human conserved non-coding sequences.* Nature, 2006. **444**(7118): p. 499-502
12. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
13. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.