Supporting Information

Benveniste et al. 10.1073/pnas.1412081111

SI Materials and Methods

Definition of Positive and Negative Sets. Promoter regions were retrieved from ENSEMBL using the annotated transcription start sites (TSSs) from the human reference genome hg19. The results shown in the paper were obtained using a window of 100 bp around the TSS; results with different window sizes are given in Table S3. In addition to TSSs, we also consider two other datasets denoted as FANTOM5 and DNase (see main text for their origin). DNase sites were downloaded from the University of California, Santa Cruz, genome browser and FANTOM5 enhancers from that project's website as BED files. These regions were defined by extending by 2 kb around the midpoint of each interval; results of the prediction of histone marks at these loci are shown in Table S4. The choice of a larger window is motivated by the greater uncertainty in defining enhancers as opposed to promoter regions. Results with a smaller window of 100 bp were similar as can be seen in Table S5. Regions were defined to be positive if they overlapped with a histone modification peak. Peaks were called from aligned reads using MACS2 (1) with false-discovery rate set to 0.01. To plot the profiles of histone marks around TSSs, mean values of reads per kilobase per million mapped reads from the two Encyclopedia of DNA Elements (ENCODE) replicates for each mark were calculated for 250-bp windows centered around each TSS.

Selection of Predictor Features. Sequence predictor features consisted of 6-mer counts within a 4-kb region centered at the TSS. The space of 4,096 possible 6-mers was further reduced to 2,080 by discarding strand information (this procedure does not exactly halve the number of features due to palindromicity of some 6-mers). Transcription factor (TF) chromatin immunoprecipitation followed by sequencing data for the three ENCODE tier 1 cell lines was downloaded from the main ENCODE repository, retaining only proteins with known transcriptional regulatory activity (filtered by gene ontology term GO:0003700—sequence-specific DNA-binding TF activity) and removing factors with known histone-modifying or chromatin-remodeling activity (e.g., EZH2 and HDAC2) to avoid including confounding factors in our analysis. TF features were associated to a TSS from aligned reads files (.BAM) by counting the number of reads mapping to within 2 kb of the TSS. Raw read counts were normalized by considering fold change with respect to the input signal. A complete list of TFs used in each cell line is given in Dataset S1.

Selection of Classification Algorithm. Logistic regression (LR) classifiers have been used throughout the paper based on two considerations. They enable direct comparison of the performance of sequence and TF-based classifiers without confounding factors relating to the algorithm used to construct these classifiers. LR-based classifiers produce easily interpretable weights that were used to verify that our method reproduces known TF–histone modifier interactions. We also show that our results are not based purely based on the choice of LR classifiers by comparing an LR classifier to a support vector machine (SVM) classifier (Fig. S2).

1. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9): R137.

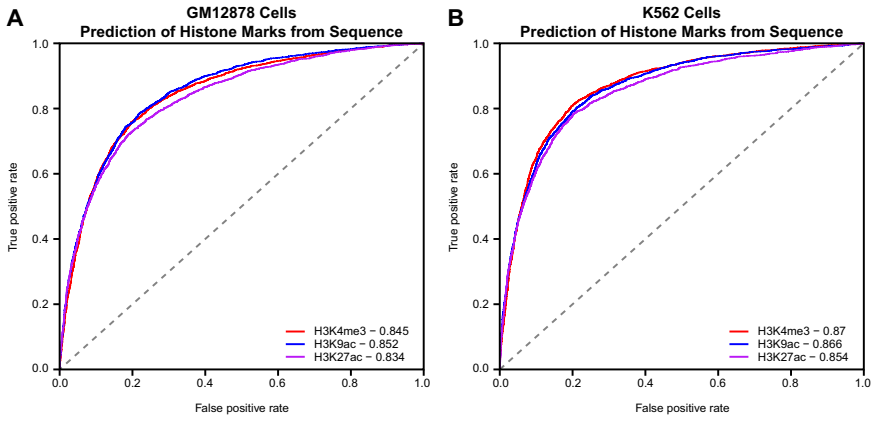

Fig. S1. Receiver operator characteristic (ROC) curves for prediction of histone modifications from DNA sequence in GM12878 (*A*) and K562 (*B*) cells. LR-based classifiers trained on a single sample of 70% of TSSs and tested on the remaining 30%. The area under curve (AUC) for each task is indicated in the legend.

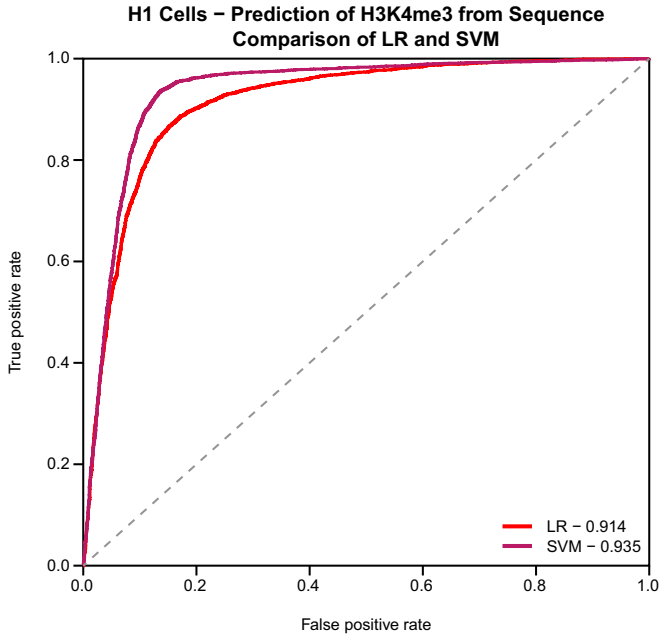

Fig. S2. Comparison of prediction of histone modifications from DNA sequence by LR and SVM. Shown are ROC curves for the prediction of H3K4me3 in H1 cells using two different classification algorithms. Classifiers were trained on the same single sample of 70% of TSSs and tested on the remaining 30%. The AUC for each task is indicated in the legend. The *k*-mer SVM was implemented and run on the Beer Lab kmer-SVM web server (1).

1. Fletez-Brant C, Lee D, McCallion AS, Beer MA (2013) kmer-SVM: A web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res* 41(Web Server issue):W544–W556.

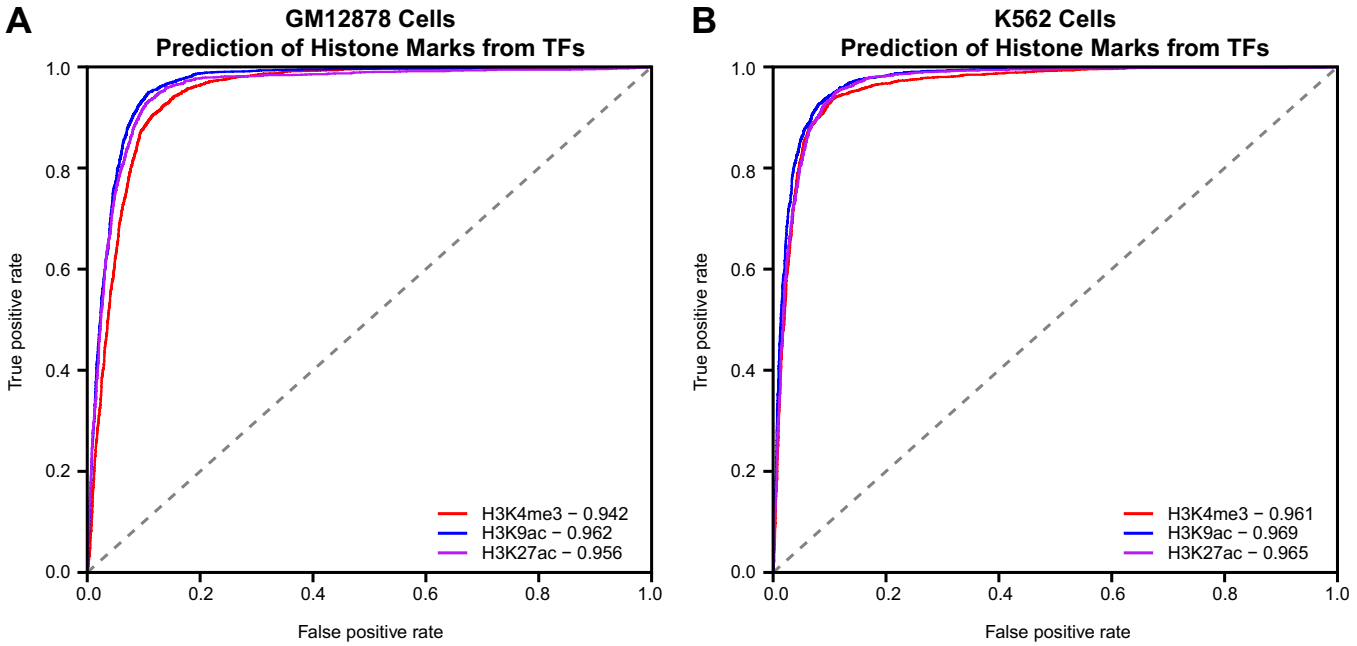

Fig. S3. ROC curves for prediction of histone modifications from TF binding in GM12878 (*A*) and K562 (*B*) cells. LR-based classifiers trained on a single sample of 70% of TSSs and tested on the remaining 30%. The AUC for each task is indicated in the legend.

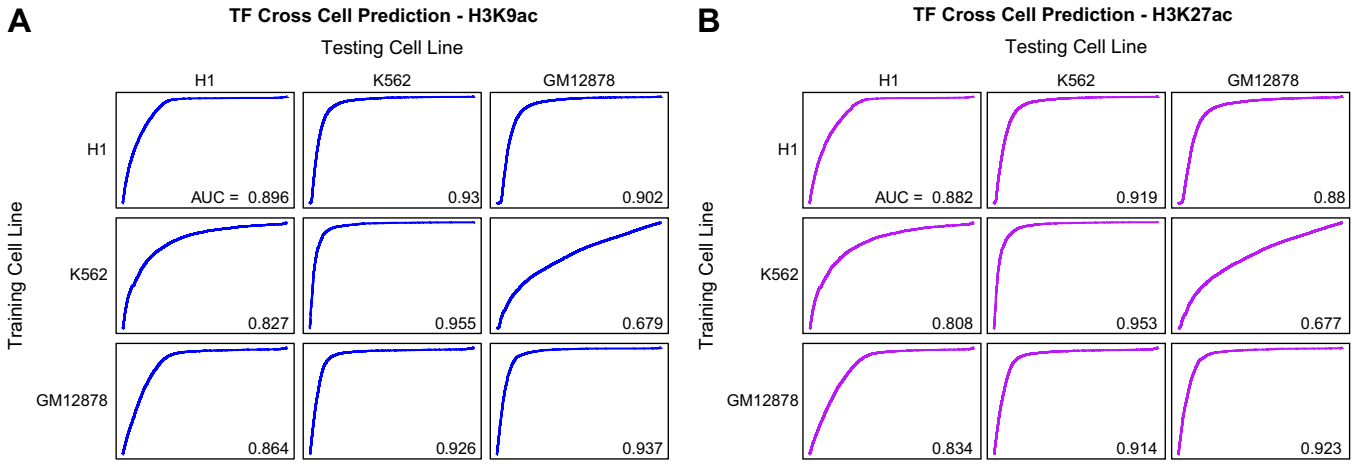

Fig. S4. ROC curves for cross-cell predictions of presence of H3K9ac (*A*) and H3K27ac (*B*). Classifiers were trained on the cell line indicated in the row and tested on each of the three cell lines (indicated by the column). The AUC is given on the plot in each case.

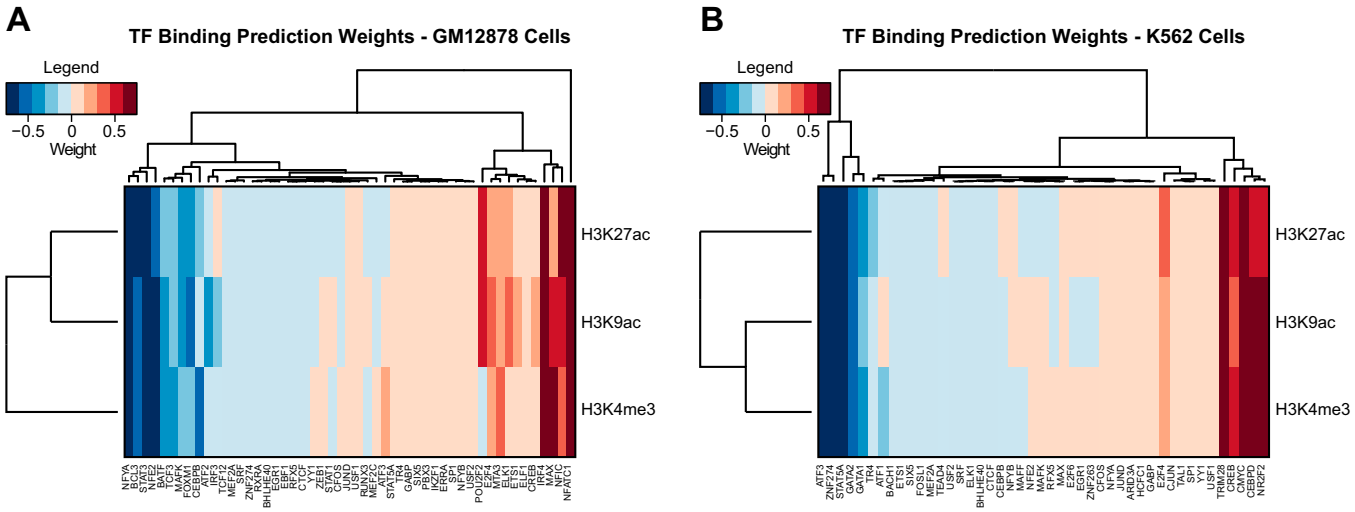

Fig. S5. Heat maps of LR weights assigned to TFs in histone mark prediction task on GM12878 (*A*) and K562 (*B*) cell lines. Each cell represents the weight assigned to a particular TF in predicting the occurrence of a particular histone mark. Both the rows and the columns were subject to hierarchical clustering.

Table S1. Degree of overlap between H3K4me modifications at human TSSs in this study

Mark	% Positive TSSs	% +ve TSSs with me3
H1 – H3K4me2	42.36% (12,635/29,828)	89.05%
GM12878 – H3K4me2	34.72% (10,356/29,828)	73.66%
K562 – H3K4me2	35.82% (10,685/29,828)	90.11%

Table S2. Accuracy measures for histone mark prediction from sequence at promoters

Prediction task	% Positive TSS	AUC (±SE)	AUPR (±SE)	Accuracy (±SE)
H1 – H3K4me3	39.8% (11,865/29,828)	0.918 (±0.001)	0.848 (±0.002)	0.856 (±0.001)
H1 – H3K9ac	24.1% (7,199/29,828)	0.867 (±0.001)	0.649 (±0.002)	0.825 (±0.001)
H1 – H3K27ac	13.8% (4,123/29,828)	0.828 (±0.002)	0.447 (±0.005)	0.861 (±0.001)
H1 – H3K27me3	8.3% (2,470/29,828)	0.806 (±0.002)	0.332 (±0.005)	0.902 (±0.001)
K562 – H3K4me3	32.6% (9,721/29,828)	0.865 (±0.001)	0.751 (±0.002)	0.814 (±0.001)
K562 – H3K9ac	29.4% (8,770/29,828)	0.865 (±0.001)	0.735 (±0.002)	0.822 (±0.001)
K562 – H3K27ac	27.6% (8,231/29,828)	0.853 (±0.001)	0.705 (±0.002)	0.821 (±0.001)
GM12878 – H3K4me3	30.7% (9,150/29,828)	0.845 (±0.001)	0.693 (±0.002)	0.800 (±0.001)
GM12878 – H3K9ac	26.2% (7,826/29,828)	0.855 (±0.001)	0.700 (±0.002)	0.814 (±0.001)
GM12878 – H3K27ac	28.1% (8,423/29,828)	0.827 (±0.001)	0.647 (±0.003)	0.806 (±0.001)

These include the area under the ROC (AUC), the area under the precision-recall curve (AUPR), and the accuracy of the logistic regression classifier at 0.5 cutoff (Accuracy). SEs are calculated from the results of 10 different train (70%)/test (30%) splits.

Table S3. Accuracy measures for prediction of histone modifications at promoters in H1 cells as the size of the window around the TSS is varied

Prediction task	100 bp	500 bp	1 kb	2 kb
Seq – H3K4me3	0.917	0.948	0.955	0.954
Seq – H3K9ac	0.864	0.893	0.900	0.898
Seq – H3K27ac	0.827	0.856	0.864	0.863
Seq – H3K27me3	0.804	0.826	0.842	0.844
TF – H3K4me3	0.952	0.976	0.982	0.983
TF – H3K9ac	0.918	0.946	0.956	0.958
TF – H3K27ac	0.908	0.927	0.934	0.935
TF – H3K27me3	0.879	0.879	0.879	0.875

Values reported are areas under the ROC curve (averages over 10 random 70/30 train/test splits). The first half of the table reports prediction performance from sequence features, and the second half from TF binding.

Table S4. Accuracy measures for prediction of histone modifications at promoters from TF binding

Prediction task	% Positive TSS	AUC (±SE)	AUPR (±SE)	Accuracy (±SE)
H1 – H3K4me3	39.8% (11,865/29,828)	0.950 (±0.001)	0.885 (±0.001)	0.8885 (±0.001)
H1 – H3K9ac	24.1% (7,199/29,828)	0.921 (±0.001)	0.729 (±0.003)	0.844 (±0.001)
H1 – H3K27ac	13.8% (4,123/29,828)	0.909 (±0.001)	0.545 (±0.003)	0.878 (±0.001)
H1 – H3K27me3	8.3% (2,470/29,828)	0.877 (±0.002)	0.437 (±0.004)	0.922 (±0.001)
K562 – H3K4me3	32.6% (9,721/29,828)	0.961 (±0.001)	0.901 (±0.002)	0.913 (±0.001)
K562 – H3K9ac	29.4% (8,770/29,828)	0.969 (±0.001)	0.906 (±0.001)	0.919 (±0.001)
K562 – H3K27ac	27.6% (8,231/29,828)	0.965 (±0.001)	0.887 (±0.003)	0.912 (±0.001)
GM12878 – H3K4me3	30.7% (9,150/29,828)	0.942 (±0.001)	0.829 (±0.002)	0.883 (±0.001)
GM12878 – H3K9ac	26.2% (7,826/29,828)	0.962 (±0.001)	0.873 (±0.002)	0.906 (±0.001)
GM12878 – H3K27ac	28.1% (8,423/29,828)	0.956 (±0.001)	0.855 (±0.002)	0.901 (±0.001)

These include the area under the ROC (AUC), the area under the precision-recall curve (AUPR), and the accuracy of the logistic regression classifier at 0.5 cutoff (Accuracy). SEs are calculated from the results of 10 different train (70%)/test (30%) splits.

Table S5. Accuracy measures for prediction from TF binding on DNase and enhancer element sets in the H1 cell line

Prediction task	% Positive TSS	AUC (±SE)	AUPR (±SE)	Accuracy (±SE)
FANTOM5 – H3K4me1	18.2% (7,825/43,011)	0.882 (±0.001)	0.626 (±0.003)	0.860 (±0.001)
FANTOM5 – H3K4me3	12.9% (5,563/43,011)	0.962 (±0.001)	0.843 (±0.003)	0.946 (±0.001)
FANTOM5 – H3K9ac	7.7% (3,297/43,011)	0.965 (±0.001)	0.736 (±0.003)	0.954 (±0.001)
FANTOM5 – H3K27ac	5.8% (2,490/43,011)	0.959 (±0.001)	0.628 (±0.004)	0.956 (±0.001)
FANTOM5 – H3K27me3	7.7% (3,307/43,011)	0.880 (±0.002)	0.530 (±0.005)	0.936 (±0.001)
DNase – H3K4me1	13.5% (172,484/1,281,988)	0.865 (±0.001)	0.498 (±0.001)	0.877 (±0.001)
DNase – H3K4me3	10.4% (133,067/1,281,988)	0.958 (±0.001)	0.821 (±0.001)	0.954 (±0.001)
DNase – H3K9ac	6.0% (77,062/1,281,988)	0.970 (±0.001)	0.762 (±0.001)	0.966 (±0.001)
DNase – H3K27ac	4.4% (56,049/1,281,988)	0.967 (±0.001)	0.642 (±0.001)	0.968 (±0.001)
DNase – H3K27me3	6.5% (83,945/1,281,988)	0.865 (±0.001)	0.456 (±0.001)	0.942 (±0.001)

These include the area under the ROC (AUC), the area under the precision-recall curve (AUPR), and the accuracy of the logistic regression classifier at 0.5 cutoff (Accuracy). SEs are calculated from the results of 10 different train (70%)/test (30%) splits. Results based on histone mark peak calls in a region of 4 kb around the enhancer midpoint.

Table S6. Accuracy measures for prediction from TF binding on DNase and enhancer element sets in the H1 cell line

Prediction task	% Positive TSS	AUC (±SE)	AUPR (±SE)	Accuracy (±SE)
FANTOM5 – H3K4me1	6.0% (2,590/43,011)	0.842 (±0.003)	0.241 (±0.004)	0.937 (±0.001)
FANTOM5 – H3K4me3	6.1% (2,626/43,011)	0.962 (±0.001)	0.587 (±0.004)	0.951 (±0.001)
FANTOM5 – H3K9ac	2.1% (885/43,011)	0.961 (±0.001)	0.344 (±0.005)	0.978 (±0.001)
FANTOM5 – H3K27ac	1.6% (679/43,011)	0.950 (±0.003)	0.306 (±0.008)	0.984 (±0.001)
FANTOM5 – H3K27me3	3.9% (1,669/43,011)	0.918 (±0.002)	0.485 (±0.006)	0.967 (±0.001)
DNase – H3K4me1	3.6% (45,751/1,281,988)	0.854 (±0.001)	0.202 (±0.001)	0.956 (±0.001)
DNase – H3K4me3	4.3% (55,289/1,281,988)	0.974 (±0.001)	0.644 (±0.001)	0.962 (±0.001)
DNase – H3K9ac	2.0% (25,233/1,281,988)	0.976 (±0.001)	0.470 (±0.001)	0.978 (±0.001)
DNase – H3K27ac	1.3% (16,033/1,281,988)	0.968 (±0.001)	0.313 (±0.001)	0.985 (±0.001)
DNase – H3K27me3	3.2% (40,990/1,281,988)	0.916 (±0.001)	0.451 (±0.002)	0.966 (±0.001)

These include the area under the ROC (AUC), the area under the precision-recall curve (AUPR), and the accuracy of the logistic regression classifier at 0.5 cutoff (Accuracy). SEs are calculated from the results of 10 different train (70%)/test (30%) splits. Results based on histone mark peak calls in a region of 200 bp around the enhancer midpoint.

Other Supporting Information Files

Dataset S1 (XLSX)