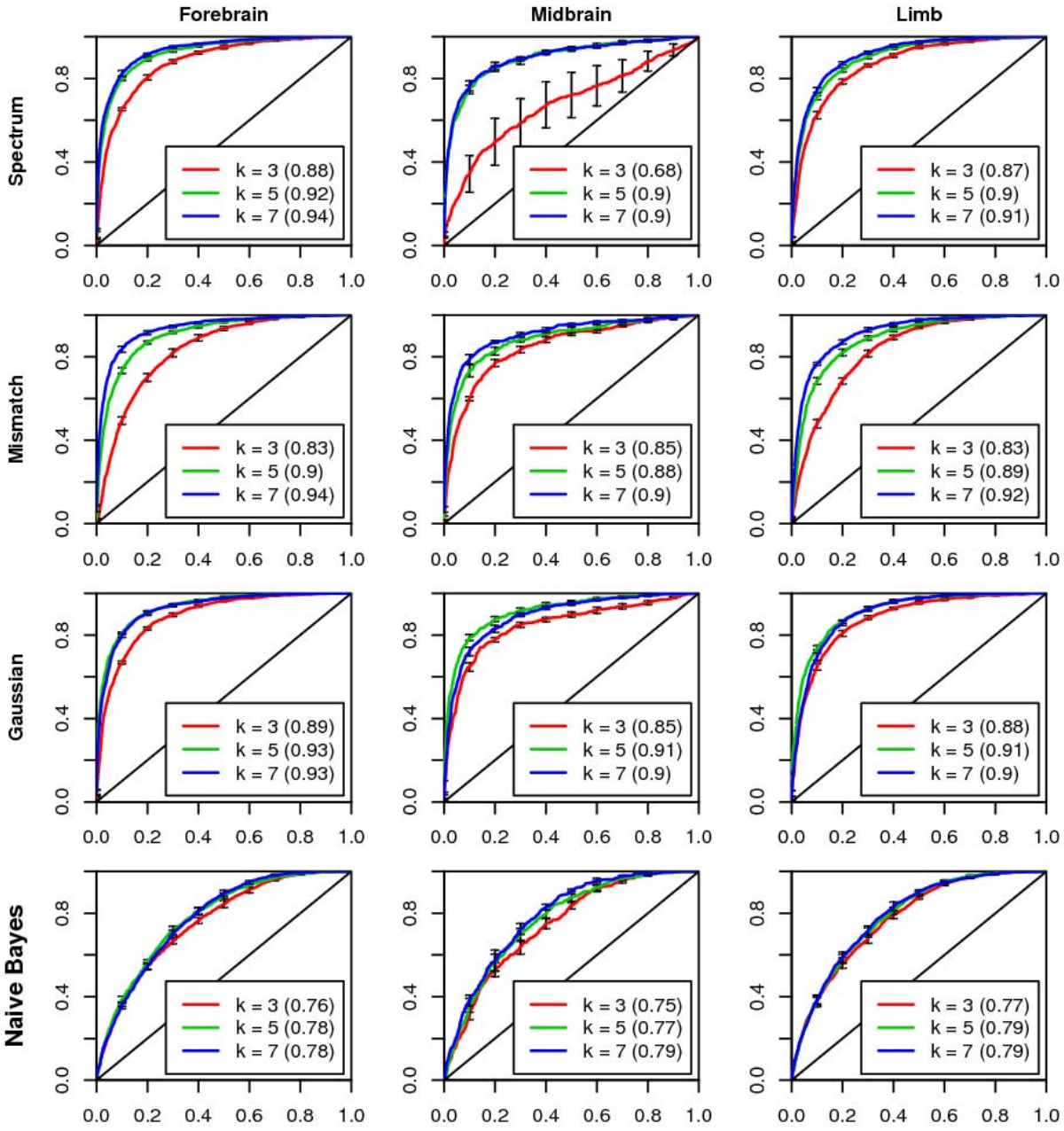
SUPPLEMENTAL MATERIAL

Supplemental Figures

Figure S1: All Classification Results of Visel's Data Set with Various Methods

(A) SVMs and Naïve Bayes Classifiers



(B) SVMs with selected 6-mers

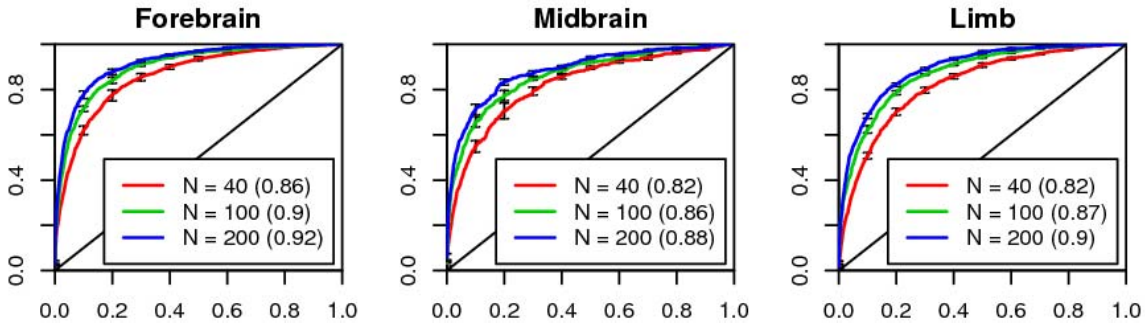


Figure S1. Here we compare the performance of SVM models with different kernels and *k*-mer lengths, and a Naïve bayes classifier. ROC curves are shown for each of the three mouse tissues. Each curve is an average of 5 cross-fold validations on a reserved test set, and error bars denote one standard deviation over the 5 cross-fold validation sets. The numbers in the parenthesis indicate the average of the area under ROC curves (auROC). Three different lengths of *k*-mers, *k*=3, 5, 7, are tested. Generally, larger *k* exhibits better performance in terms of auROCs with some exceptions caused by over-fitting. (A) Using the full set of *k*-mers, SVM Classification results with three different kernels (Spectrum, Mismatch, and Gaussian) and Naïve Bayes classification results are shown. SVMs outperform Naïve Bayes classifiers in every case but one which failed to converge (SVM with 3-spectrum kernel on Midbrain). (B) Using only selected 6-mers, results of SVMs with spectrum kernels are presented. For each classification, a half of *N* 6-mers with the largest positive SVM weights and a half of *N* 6-mers with the largest negative SVM weights were selected (*N*=40, 100 and 200).

Figure S2: Length Distribution and Repeat Fraction Distribution Between Enhancers and Random Genomic Sequences Matched to EP300 enhancer set

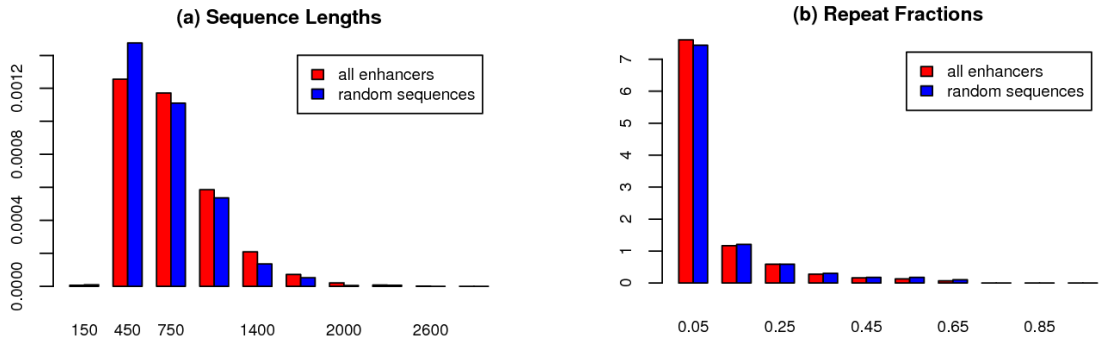


Figure S2. For our null-sequence model, we selected random sequences from the genome to match the repeat fraction and length distribution of the sequences in the EP300 data set. The combined set of all Visel's EP300 bound regions are shown in red, and our null sequence set is shown in blue.

Figure S3: Comprison between ROC curves and Precision-recall curves with larger negative sets

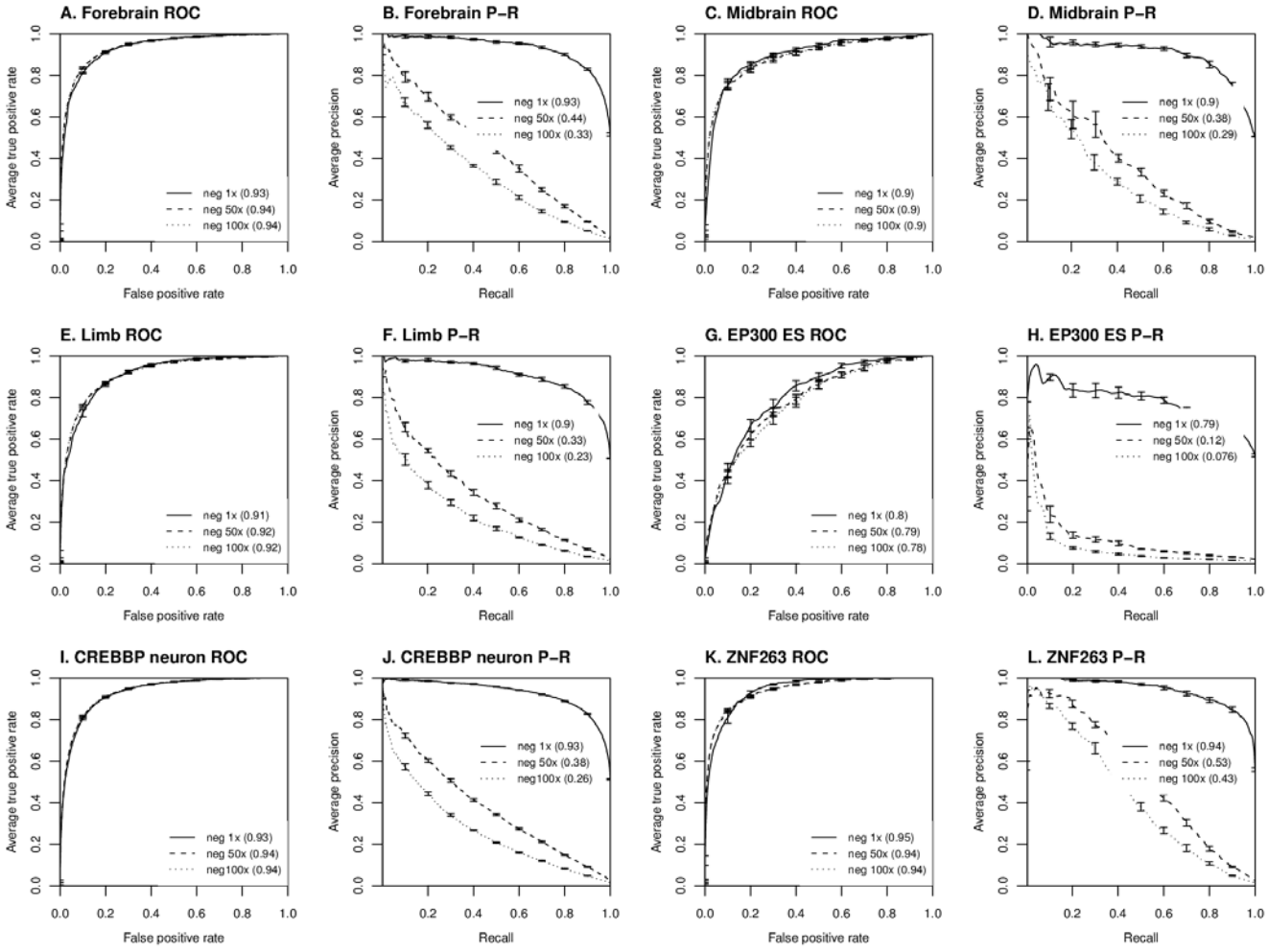


Figure S3. Here we compare the scaling of negative set size for all comparisons of positive sets vs. random genomic sequence for the 6-mer spectrum kernel SVM (see Table S1). The genomic ratio of enhancers to non-enhancer sequence is very large (we estimate that enhancers comprise 1-2% of the genome), so we used three negative sets (1x; 50x larger; and 100x larger than the positive enhancer set) for each case. The area under the ROC curve (auROC) or the area under the precision-recall curve (auPRC) is shown in parentheses. For large negative set size, auPRC is a more reliable measure of performance than the auROC curve, which is independent of negative set size, as expected.

Figure S4: Comparison Between Frequencies and SVM Weights of *k*-mers

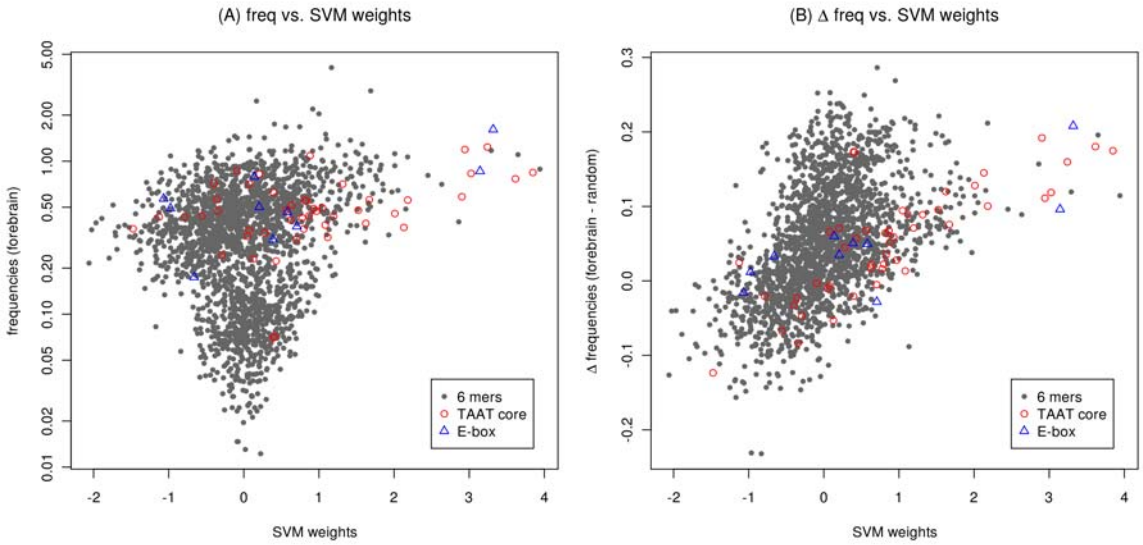


Figure S4. While the SVM features which are assigned large positive weights are generally over-represented in the EP300 bound regions relative to background genomic sequence, there is not a strictly direct correlation between SVM weights and *k*-mer frequencies. (A) *k*-mer frequency in forebrain vs. SVM weights. (B) Normalized frequency difference between forebrain and random sequences, $\Delta f = (freq(fb)-freq(rand))/(freq(fb)+freq(rand))/2$.

Figure S5: Average EP300 ChIPseq Read Coverage in the SVM Predicted Regions

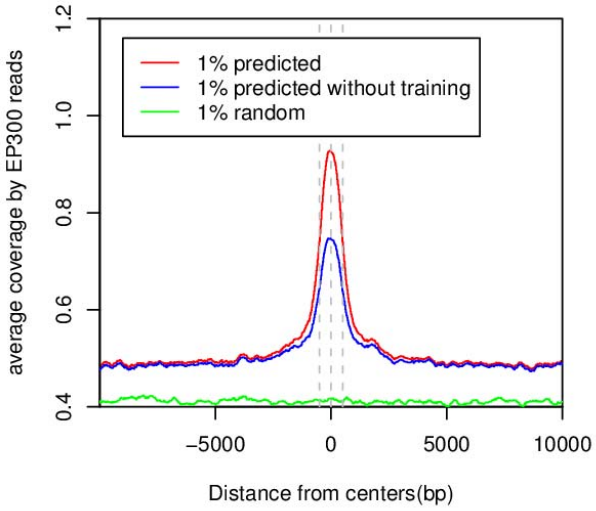


Figure S5. EP300 reads are significantly enriched in the SVM predicted regions: The middle point of the top 1% SVM predicted regions in forebrain were aligned at 0bp, the sequence around each peak was extended +-10kb in each direction, and the average coverage of EP300 reads in the surrounding regions is shown. Significant enrichments compared to random genomic sequence (by about two fold) is observed even after those regions which overlap with the original training set are excluded. This is further evidence that the SVM predicted regions which are not in the EP300 positive test set are in fact EP300-bound.

Figure S6: Correlation of SVM predictions and EP300 read density for genome wide scan.

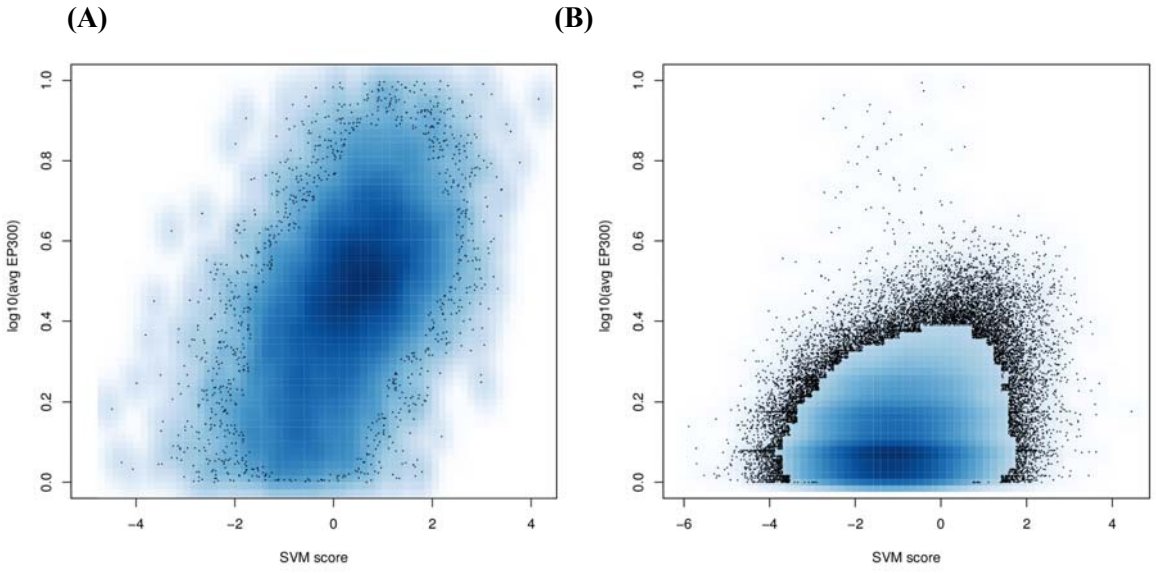


Figure S6. Here we show the correlation of our SVM score and EP300 read density for all 1kbp regions across the genome. On the left (A) are all regions that partially overlap any positive training set region. On the right (B) are all other genomic regions. The cloud of points with EP300>2 ($\log_{10}2$=0.301) and SVM score>1 are our predicted enhancers, and we expect about 50% of these to be true positive enhancers. Most regions with EP300>3 are be in the positive training set, and are in (A) by construction. The regions in (A) with EP300<3 are genomic 1kbp chunks which partially overlap a positive training set region.

Figure S7: Distribution of SVM scores for varying negative set size.

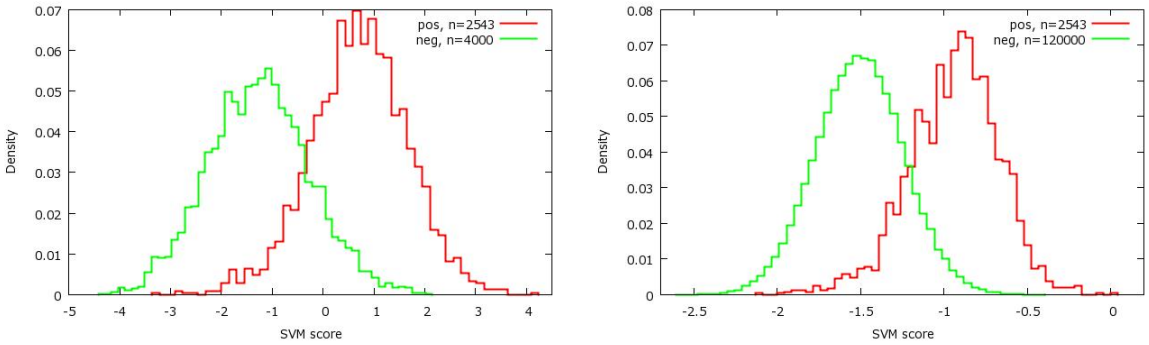


Figure S7. The distributions of SVM scores for negative set size (N=4000) and (N=120,000) are shown. While there is a shift in the (arbritary) scale, the distributions are very similar, reflecting the fact that auROC is similar for N=4000, N=120,000, or N=240,000 negative sequences. On the other hand, as the negative set size increases, auPRC drops, because the higher scoring tail of the negative sequence score distribution becomes comparable to the bulk distribution of the positive sequences.

Figure S8: Correlation between SVM scores from two separately trained SVMs

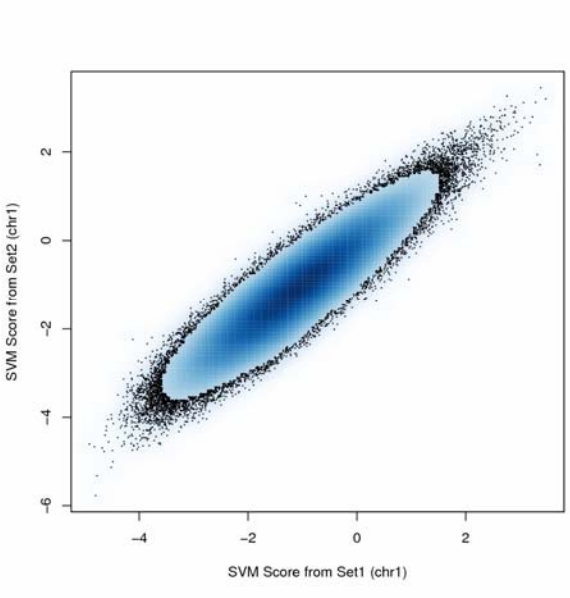


Figure S8. To investigate the robustness of the top SVM scoring regions, we trained separate SVMs using independently sampled random negative sequence sets, and compared the top SVM scoring regions using these different negative sequence sets. While there is some variation between the top scoring regions from different negative sets, only rarely do high scoring regions in one SVM not score highly the other SVMs, indicating that the predictions are robust to different realizations of the negative set. As shown in Table S2, there is 64.5% overlap between the top 1.0% regions for "Set1" and "Set2" SVMs, but 84.5% and 92.2% of the top 1% sites in Set1 are found in the top 2% and 3% of Set2, respectively. Figure S8 compares the scores of chromosome 1 regions (to reduce the number of plotted points) from these two SVMs, showing very high correlation (C=0.915).

Figure S9: Classification of Human Homologous Regions of the EP300 Mouse Training Set

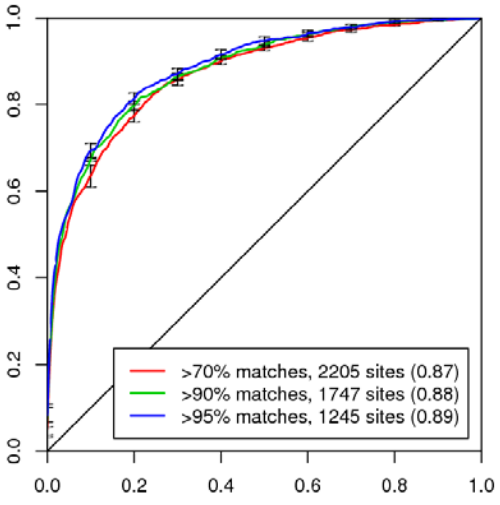


Figure S9. SVMs can discriminate human homologous EP300 bound regions from human random sequence. A positive human test set was generated by sequence alignment of the mouse EP300 training set regions to the human genome, varying the stringency for assigning homologous regions (70% identical, 90% identical, and 95% identical). As shown in Figure S9, all three of these sets can be classified with high accuracy (auROC=0.87, 0.88, 0.89), and classification power is relatively unaffected by the cut-off for determining homologous regions, again demonstrating the robustness of our SVM predicted enhancers.

Figure S10: SVM preditions at the human *Otx2* locus.

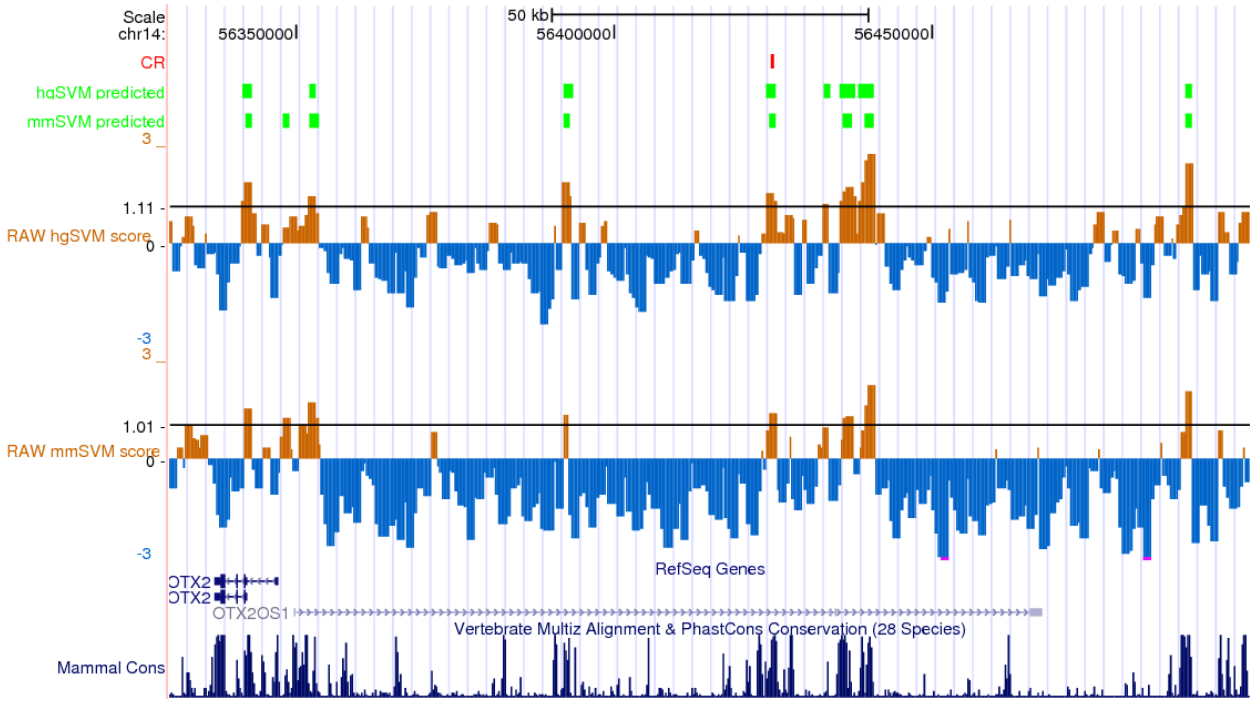


Figure S10. To further compare the predictions of the SVM trained on the mouse EP300 bound regions and the SVM trained on human homologous sequence, we used these two SVMs (mmSVM and hgSVM) to score the human genome *Otx2*, which is known to play a role in forebrain development. The raw hgSVM and mmSVM scores are quite similar, and most of the predicted enhancers above the 1% threshold overlap. One of these enhancers has been experimentally verified to have enhancer activity (CR).

Figure S11: 6-mer SVM scores across the SOX2-POU5F1(OCT4)-NANOG binding consensus:

	SOX2 OCT4	weights
SOX2-OCT4-NANOG:	**CATTGTYATGCAAAT**	
SOX2:	**CATTGT**`.........`	1.45
SOX2:	`.`**ATTGTY**`........`	1.19 or 0.77
SOX2:	`..`**TTGTYA**`.......`	0.85 or 0.32
OCT4:	`......`**TATGCA**`...`	1.00
OCT4:	`.......`**ATGCAA**`..`	0.71
OCT4:	`........`**TGCAAA**`.`	0.58
OCT4:	`.........`**GCAAAT**	0.47

Figure S11. Many large weight *k*-mers from the SVM trained on the EP300 ES dataset are subsequences that tile across the SOX2-OCT4 consensus oligo.

Figure S12: PWM vs. *k*-mers as feature sets on forebrain and ZNF263

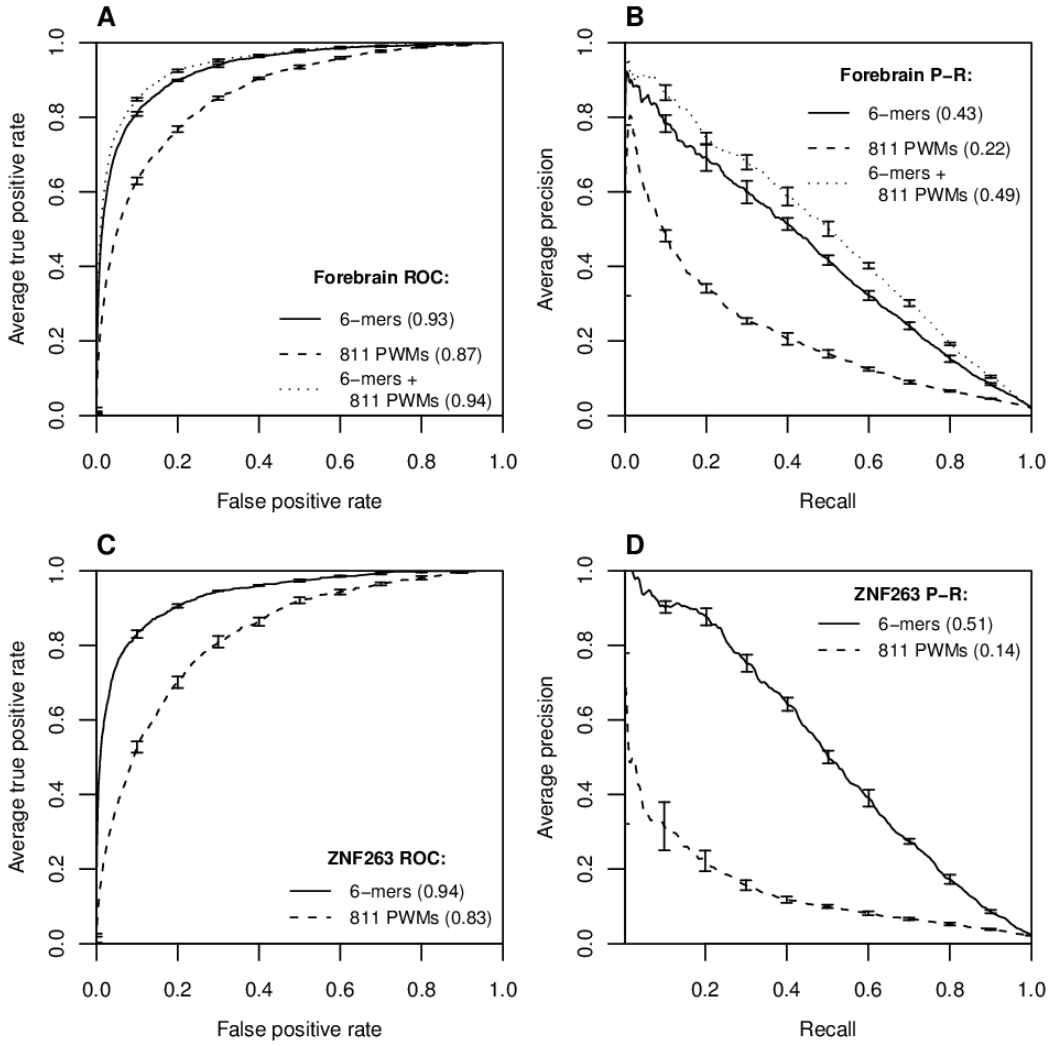


Figure S12. Here we compare SVM performance using *k*-mers to an SVM using 811 known PWMs as features using ROC (A,C) and P-R curves (B,D). (A) On the forebrain enhancers, our *k*-mer SVM is more accurate than known PWMs alone, but a combination of *k*-mers and PWMs performs slightly better. (B) These differences in auROC translate to a dramatic reduction in auPRC for PWMs relative to *k*-mers only or combined *k*-mers and PWMs. (C) Our *k*-mer SVM predicts ZNF263 bound regions from ChIP-seq with high accuracy (auROC=0.94), but the 811 PWM SVM is less accurate (auROC=0.83). (D) Again the lower auROC for PWMs corresponds to a significant decrease in auPRC for PWMs on the ZNF263 data (0.14 vs. 0.51), and a much higher false discovery rate.

Figure S13: Classifications Using One Negative Set Shared Between Different Datasets

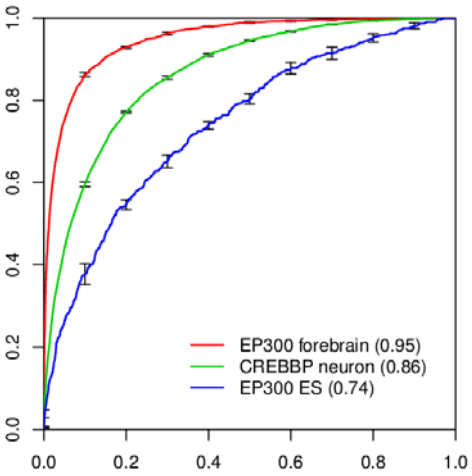


Figure S13. When training the SVMs for the three data sets (EP300 forebrain, CREBBP neuron, and EP300 ES), we used independent negative sets. To ensure that the predictive *k*-mers with large negative weights reflect their absence in the positive training set, not presence in various negative set realizations, we also generated one common negative set shared between the three data sets. Since the length distribution and repeat fractions of the three data sets are different, we modified the length of the positive sets to be able to generate a single appropriate negative set. For Chen's and Kim's dataset, we extended a fixed length from the peaks reported. We chose 800bps (+-400bp from the peaks) to match with the lengths of forebrain data set as closely as possible (mean length of the forebrain data set is 816bp). We also chose the fixed 800bp length for the negative set because forebrain data set was relatively unaffected by the length distribution. We then sampled 20000 random genomic sites for the negative set. To deal with the unbalanced positives and negative set sizes, we optimized class weights for the positive sequences, and report the best result of each case. Figure S13 shows the ROC curves of three different dataset classifications against the common negative set. This result is comparable to the original analysis (Figure 2G). Table S8 shows the top 15 positive 6-mers and top 10 negative 6-mers of each dataset from this analysis, which largely overlap the results from the independent random negative sets, as shown in Tables 1, S5 and S6.

Figure S14: auROC and BEP using single chromosomes as test sets in 20-fold cross validation.

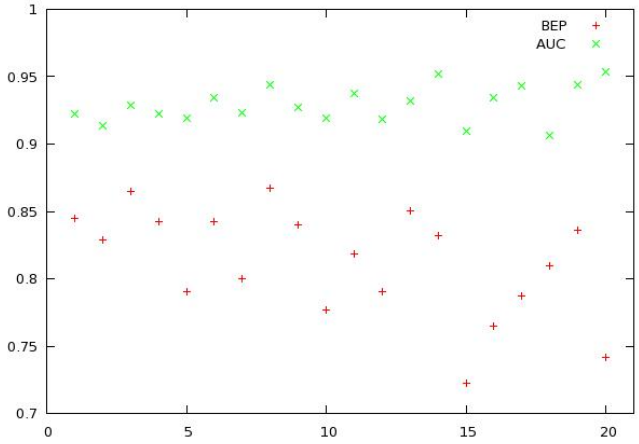


Figure S13. To show that our cross validation procedure does not impact classification performance, we also trained SVMs using 20-fold cross validation with test sets consisiting of all elements on a single chromosome (1-19, and X=20), instead of 5-fold cross validation used in the main text. The variation in auROC and Precision at the break-even point (BEP, where precision equals recall) is consistent with the varying size of the test sets. No chromosome is significantly more or less accurately predicted than the others.

Supplemental Tables

Table S1: Outline of Analysis

Classification Experiments					
Experiment	Positive	Negative	Neg. set size	Applied Methods	Figures
Visel EP300 vs random	EP300 fb, mb, and lb	random genomic	4000	sk-SVM, msk-SVM, G-SVM, NB, L1-LR*, and KIRMES*	Fig 2, Fig S1
Visel EP300 vs each other	EP300 fb,	EP300 mb	-	sk-SVM	
	EP300 fb	EP300 lb			
	EP300 mb	EP300 lb			
Negative set size scaling	EP300 fb, mb lb, EP300 ES, CREBBP neuron, ZNF263	random genomic	1x,50x,100x	sk-SVM	Fig S3
Human	EP300 fb human orthologs	random human genomic	1x	sk-SVM	Fig S9
Other EP300/CREBBP datasets vs random	EP300 ES	random genomic	10x	sk-SVM	Fig 7
	CREBBP neuron	random genomic	1x	sk-SVM, WDS	
Other datasets vs each other	fb	EP300 ES	-	sk-SVM	
	fb	CREBBP neuron			
	EP300 ES	CREBBP neuron			
TF ChIP-seq	ZNF263	random genomic	50x	sk-SVM	Fig S12
k-mers vs PWM	EP300 fb, ZNF263	random genomic	4000, 50x	811 PWM SVM	Fig S12
Assessment Methodologies					
Similarity between predictive *k*-mers and known TFBS					Table 2,4,5
Conservation of predictive features between mouse and human					Fig 3
Spatial constraints between predictive features and relevant genes					Fig 4,6
DNaseI hypersensitivity of genome-wide enhancer predictions					Fig 5

* only applies to EP300 fb vs random genomic; sk-SVM, spectrum kernel; msk-SVM, mismatch spectrum kernel; G-SVM, Gaussian kernel; NB, Naïve Bayes; L1-LR, logistic regression; WDS, weighted degree kernel with shifts.

Table S2: Overlap Between Top SVM Scoring Regions Determined by Two Separately Trained SVMs

Set1	Set2	Number of sites only in set1 (a)	Number of sites only in set2	Number of sites in both sets (b)	% Overlap (100*b/(a+b))
0.5%	0.5%	9815	9815	16575	62.8
1.0%	1.0%	18711	18711	34069	64.5
1.0%	2.0%	8156	60938	44624	84.5
1.0%	3.0%	4119	109679	48661	92.2
1.0%	5.0%	1313	212432	51467	97.5

Table S3: Human Enhancer Prediction Using a Mouse vs. a Human SVM

Set1 (mouse)	Set2 (human)	Number of sites only in Set1 (a)	Number of sites only in Set2	Number of sites in both sets (b)	Overlap (100*b/(a+b))
0.5%	0.5%	19147	19147	9531	33.2
1.0%	1.0%	36819	36818	20537	35.8
1.0%	2.0%	28489	85843	28867	50.3
1.0%	3.0%	23352	138060	34004	59.3
1.0%	5.0%	16825	246242	40531	70.7

To further quantify the similarity of the predictions from the mouse and human SVMs, Table S2 shows the overlap of the top SVM scoring regions of the two SVMs. The mouse SVM (Set1) uses the mouse EP300 training set as positives and mouse random genomic regions as negatives, and the human SVM (Set2) uses human homologous regions of the mouse EP300 training set as positives and human random genomic regions as negatives. One third of top 1% scoring regions of Set1 are also found in the top 1% scoring regions of Set2. This overlap is quite significant considering the fact that the two SVMs were learned on different genomes.

Table S4: Overlap Between EP300 and CREBBP binding in different Data Sets

(a) EP300-bound regions in each tissue of mouse embryo vs. CREBBP peaks in activated cultured neurons

Regions	Number of observed peaks in the regions	Number of total peaks	Probability of peaks being in a region	Expected number of peaks in the regions	*p*-value
forebrain	408	11847	0.021	249.64	< 2.2e-16
midbrain	72	11847	0.004	50.18	0.002143
limb	67	11847	0.015	179.27	1

(b) EP300 bound regions in each tissue of mouse embryo vs. EP300 peaks in embryonic stem cells

Regions	Number of observed peaks in the regions	Number of total peaks	Probability of peaks being in a region	Expected number of peaks in the regions	*p*-value
forebrain	11	524	0.021	11.04	0.5464
midbrain	3	524	0.004	2.22	0.3826
limb	1	524	0.015	7.93	0.9997

Here we assess the significance of the overlap between Visel's EP300 bound regions and two other data sets: EP300 bound regions in ES cells and CREBBP bound regions in activated neurons. We count the number of EP300/CREBBP ChIP-seq peaks in the new data sets which are located within the regions of Visel's data set (EP300 forebrain, midbrain, and limb), and calculate the *p*-value of the overlap. For a null hypothesis, we assume that the observed peaks could have been detected anywhere in potential regulatory regions, which we conservatively estimate as roughly 3.5% of entire genome (Waterston et al. 2002). Then the *p*-value of the overlap is calculated from the binomial distribution.

Table S5 – Predictive 6-mers of CREBBP Neuron

(A) Fifteen 6-mers with the largest positive SVM weights

6-mers	Reverse Complement	SVM weight	Database Family Match	Top Matched Transcription Factors (q-val <0.1)
GACTCA	TGAGTC	5.68	Leucine Zipper	BACH1, NFE2, BACH2, JUNDM2, AP1
CAGATG	CATCTG	5.25	HLH	ZNF238, TAL1:TCF3, TAL1:TCF4, TCF3
GAGTCA	TGACTC	5.19	Leucine Zipper	BACH1, BACH2, NFE2, JUNDM2, AP1
ACGTCA	TGACGT	5.05	Leucine Zipper	ATF6, CREB, ATF1
TGCCAA	TTGGCA	4.80	Nuclear Factor I	HIC1, NFIC, NF1
AGATGG	CCATCT	4.29	HLH	TAL1:TCF3, TAL1:TCF4, YY1, TCF3
CATATG	CATATG	4.13	-	-
AATTAG	CTAATT	4.02	Homeodomain	VSX2, PRRX2, EVX2, PDX1, GBX2
GGCAAC	GTTGCC	3.63	-	-
CTGGCA	TGCCAG	3.47	Nuclear Factor I	HAND1::TCF3, HIC1, NFIC, TGIF2
TAATTA	TAATTA	3.42	Homeodomain	OTP, PROP1, HOXA, ALX1, LHX3
GATTCA	TGAATC	3.30	-	-
CGTCAC	GTGACG	3.18	Leucine Zipper	PAX3, CREB, ATF1, JUNDM2
GCGTCA	TGACGC	3.12	Leucine Zipper	CREB, ATF6, BACH1, BACH2
AATTAC	GTAATT	3.10	Homeodomain	PRRX2, HOXA6, HOXA1, HOXC8, DLX1

(B) Five 6-mers with the largest negative SVM weights

6-mers	Reverse Complement	SVM weight	Database Family Match	Top Matched Transcription Factors (q-val <0.1)
GGTCAA	TTGACC	-3.58	Nuclear Receptor	PPARG, RORA2, HNF4A, RORA1, ESRRA
CTGACC	GGTCAG	-3.98	Nuclear Receptor	RORA2, RORA1, NR2F2, ESRRA, PPARG
AGGTGA	TCACCT	-3.99	Zinc-finger	ZEB1
CAGGTA	TACCTG	-4.08	Zinc-finger	ZEB1
GGGTCA	TGACCC	-4.64	Nuclear Receptor	NR2F2, ESRRA, HNF4A, RXRA, PPARG

Table S6 – Predictive 6-mers of embryonic stem cells

(A) Fifteen 6-mers with the largest positive SVM weights

6-mers	Reverse Complement	SVM weight	Database Family Match	Top Matched Transcription Factors (q-val <0.1)
ACAATG	CATTGT	1.45	SOX	SOX17, SOX9, SOX5, SOX10, SOX30
ATTGTC	GACAAT	1.19	SOX	SOX17
ACAAAG	CTTTGT	1.06	SOX	SOX4, SOX11, SOX10, HNF4
TATGCA	TGCATA	1.00	Homeodomain	POU2F1, POU3F3, POU2F3, POU2F2
GAGCTA	TAGCTC	0.95	-	-
CAAAAG	CTTTTG	0.90	-	-
AGGTCA	TGACCT	0.89	Nuclear Receptor	RORA1, PPARG, RORA2, ESRRB, RAR
AAAGCC	GGCTTT	0.89	-	-
AATTCC	GGAATT	0.88	-	-
AAGGTC	GACCTT	0.88	Nuclear Receptor	PPARG, ESRRB, ESRRA, RAR, NR2F2
TCTACA	TGTAGA	0.85	-	-
TAACAA	TTGTTA	0.85	SOX	SOX5
CCGGAA	TTCCGG	0.84	ETS	ELK4, GABPA, NRF2, STAT3, ELK1
GGTGAC	GTCACC	0.82	Leucine Zipper	SREBP, CREB, AP1, ATF, RAR
CATTCA	TGAATG	0.79	SOX	HBP1

(B) Five 6-mers with the largest negative SVM weights

6-mers	Reverse Complement	SVM weight	Database Family Match	Top Matched Transcription Factors (q-val <0.1)
AACATG	CATGTT	-0.75	-	-
GCTAGA	TCTAGC	-0.76	-	-
CTGATA	TATCAG	-0.82	-	-
AATAAA	TTTATT	-0.83	Homeodomain	HOXD13, FOXC1, HOXB13
ACAAAT	ATTTGT	-0.85	-	-

Table S7: Comparison of Predictive *k*-mers from the Different Data Sets

(a) Fifteen 6-mers with the largest positive SVM weights

ES (+) vs forebrain (-)		ES (+) vs neuron (-)		forebrain (+) vs. neuron (-)	
AGGTCA	2.83917	AGGTCA	1.38178	ACAAAG	3.20456
ACCTTG	2.34974	CAATAG	1.08993	AGCTGC	2.63237
AGGTGA	1.97262	ACCTTG	0.954087	TAATGA	2.4704
CCTTGA	1.77742	AAGGTC	0.853491	CAGCTG	2.4703
AAGGTG	1.76077	CCTTGA	0.760972	GAACAA	2.46156
CACACC	1.64343	GGTCAC	0.73022	AAAGGG	2.33583
GGTGGA	1.52651	CCGGAG	0.709136	GGATTA	2.23093
CAGGTA	1.51114	ACCTGA	0.671658	ACAATG	2.22657
CACCTG	1.48868	CACCTG	0.666395	CAATTA	2.22224
CTGACC	1.45141	CAGGTA	0.639238	AATTAG	2.10941
ACCTGG	1.44535	TCTACA	0.628242	CAATGG	2.08748
AGGTAA	1.42999	GGTCAA	0.627209	ATTAGC	2.05797
GAGTCA	1.38264	ACACCC	0.624997	GGCCCC	2.01011
CTAGAA	1.32414	GAACCC	0.624728	ACAATA	1.89593
AGGAAG	1.30616	AGGTGA	0.619084	GAGGCC	1.88978

(b) Fifteen 6-mers with the largest negative SVM weights

ES (+) vs forebrain (-)		ES (+) vs neuron (-)		forebrain (+) vs. neuron (-)	
CTGGCA	-1.36987	ACAGAT	-0.58506	ATTTCA	-1.49916
AGGGGG	-1.38789	ATGACG	-0.593	GTGCCA	-1.52618
AATTAG	-1.39198	AACATG	-0.605551	CTCATC	-1.53853
GCTGCC	-1.41468	ATGCCA	-0.629451	ACTCAT	-1.60758
CAGCTG	-1.49427	AATATG	-0.637472	ACGTCA	-1.67551
CAGATG	-1.5308	CTAAAA	-0.647749	AACATG	-1.69289
ACAAAG	-1.54907	TAATTA	-0.660014	AAATTA	-1.71518
TAATTA	-1.6466	GGCAAC	-0.660122	GTTCCA	-1.73856
AGCTGC	-1.65257	CAGATG	-0.670137	GACTCA	-1.77009
GGCAAC	-1.66172	ACGTCA	-0.726237	CTAAAA	-1.9048
CAATTA	-1.68426	TGCCAA	-0.773441	GAATCA	-1.90671
AATGAG	-1.68725	CTGGCA	-0.776692	CATCTC	-1.94972
AATTAA	-1.76872	ATGTCA	-0.817955	GATTCA	-2.34549
CACAAA	-1.80749	GAAATA	-0.824868	GAGTCA	-2.46636
TAATGA	-1.882	AAATAG	-0.833908	TGCCAA	-3.12478

Table S8: Predictive k-mers of three different datasets using common random negative sequences.

(a) Fifteen 6-mers with the largest positive weights

Forebrain		Neuron		ES (w=5)	
AATTAG	5.166	TGCCAA	6.09624	AGGTCA	4.11367
CAATTA	4.53013	GACTCA	5.72339	ACAATG	3.33138
CAGCTG	4.00637	GAGTCA	4.82888	ACCTTG	2.98221
AGCTGC	3.95227	CAGATG	4.59956	AAGGTC	2.81142
ACAAAG	3.91098	CATATG	4.39705	CAATAG	2.66733
TAATGA	3.82868	AGATGG	4.28309	ATTGTC	2.44483
ATTAGC	3.79013	CTGGCA	4.23722	AAGCCG	2.24401
AATGAG	3.70752	ACGTCA	4.17313	CCGGAG	2.20552
TAATTA	3.60443	AATTAG	4.01272	CCGCCA	2.11105
CAGATG	3.49841	TAATTA	3.74813	TCGGAA	2.08395
GGCAAC	3.30678	GGCAAC	3.51165	CGGCTA	2.00939
ACAATG	2.79336	AATTAC	3.49265	ACGCCT	1.99696
CATTCA	2.71148	GAATCA	3.33674	ACACCC	1.9699
TGCTAA	2.52389	CTTGGC	3.30323	AATCCC	1.9580
AATTGG	2.43143	GCCAAA	3.23247	AATTCC	1.95536

(b) Ten 6-mers with the largest negative weights

Forebrain		Neuron		ES (w=5)	
AGGTAA	-1.71708	AGGTCA	-2.30776	GTCATA	-1.69074
CCATGA	-1.73288	CGGTCA	-2.3817	ATGTCA	-1.71948
AGGATG	-1.73706	AAGGTG	-2.46491	CACTAC	-1.75958
ACCTGA	-1.89247	AAAGTT	-2.54574	AACATG	-1.78623
AGGTAG	-1.99038	GGTCAA	-2.65552	ATCCAA	-1.78928
AGGTGA	-2.03761	ATGACC	-2.66935	AAAAAG	-1.87093
CATCTA	-2.10399	TGACCA	-2.76447	GAAATA	-1.94475
AAGTCA	-2.1352	AGGTAA	-2.99852	GACCGA	-2.02932
ACCTGG	-2.39139	AGGTGA	-3.39747	ACTCGA	-2.12442
CAGGTA	-2.92076	CAGGTA	-4.02767	CAATTC	-2.19647

Supplemental Methods

Naïve Bayes Classifier

To compare our SVM to an alternative approach, our implementation of the Naïve Bayes classifier follows (Yuan et al. 2007). The Naïve Bayes classifier calculates the posterior probability of the class of each sequence. The Naïve Bayes classifier uses the same full set of *k*-mers, and converts them into binary vectors using a threshold that maximizes χ^2 for each *k*-mer independently. A *k*-mer with frequency above the maximum χ^2 threshold is "present" in that test sequence. Then assuming the conditional independence between features given a class, the posterior probability simplifies via Bayes rule to:

$$P(Y \mid \mathbf{X} = \mathbf{x}_i) = \frac{P(\mathbf{X} = \mathbf{x}_i \mid Y)P(Y)}{P(\mathbf{X} = \mathbf{x}_i)} = \frac{P(Y)}{P(\mathbf{X} = \mathbf{x}_i)} \prod_j P(X_j = x_{i,j} \mid Y)$$

The ratios between $P(Y=1|\mathbf{X}=\mathbf{x}_i)$ and $P(Y=0|\mathbf{X}=\mathbf{x}_i)$ is finally used as a score to classify each test sequence.

PhastCons Conservation Score

To measure conservation, we use PhastCons (Siepel et al. 2005), based on a two-state phylogenetic hidden Markov model (phylo-HMM). PhastCons outputs the probability that each aligned column was generated by the "conserved" state given the model parameters and the multiple alignment. We used this PhastCons conservation score for alignment of 29 vertebrate genomes available at the UCSC Genome Browser (Karolchik et al. 2008) to assess how well the predictive sequence elements in the enhancers are evolutionarily conserved. We calculated the average PhastCons score over all bases of each *k*-mer in each sequence, and obtained one score for each *k*-mer ranging from 0 to 1, reflecting overall conservation.

Supplemental References

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H., Diekhans M, Giardine B, Harte RA, Hinrichs AS., Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucl Acids Res* **36**: D773-779.

Siepel A, Bejerano G, Pedersen Jakob S., Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034 -1050.

Yuan Y, Guo L, Shen L, and Liu JS. 2007. Predicting Gene Expression from Sequence: A Reexamination. *PLoS Comput Biol* **3**: e243.

Supplemental Data

1. Sequence Files in FASTA Format (sequence_file.tar.gz)

Visel's data set

EP300 forebain: enh_fb.fa

EP300 midbrain: enh_mb.fa

EP300 limb: enh_lb.fa

Random genomic sequences: random4000.fa

Kim's data set

CREBBP neuron (+-100bp): enh_neuact_200.fa

CREBBP neuron (+-400bp): enh_neuact_800.fa

Random genomic sequences: random12000.fa

Chen's data set

EP300 ES (+- 100bp): enh_es_200.fa

EP300 ES (+- 400bp): enh_es_800.fa

Random genomic sequences: random5240.fa

2. SVM Weight File (svmweights.xls)

EP300 forebrain (+) vs random (-): fb_vs_rnd

EP300 midbrain (+) vs random (-): mb_vs_rnd

EP300 limb (+) vs random (-): lb_vs_rand

CREBBP neuron (+) vs random (-): neu_vs_rand

EP300 ES (+) vs random (-): es_vs_rand

forebrain (+) vs limb (-): fb_vs_lb

midbrain (+) vs limb (-): mb_vs_lb

forebrain vs midbrain: NA

EP300 ES (+) vs EP300 forebrain (-): es_vs_fb

EP300 ES (+) vs CREBBP neuron (-): es_vs_neu

EP300 forebrain (+) vs CREBBP neuron (-): fb_vs_neu

3. forebrain enhancer predicted regions with SVM>=1.0 (enh_fb_pred.xls)