Discriminative prediction of mammalian enhancers from DNA sequence

Dongwon Lee, Rachel Karchin and Michael A. Beer

Genome Res. published online August 29, 2011
Access the most recent version at doi:10.1101/gr.121905.111

Discriminative Prediction of Mammalian Enhancers from DNA Sequence

Dongwon Lee[1], Rachel Karchin[1,2] and Michael A. Beer[1,3]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA

[2]Institute for Computational Medicine, Johns Hopkins University, Baltimore, USA

[3]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, USA

Corresponding author, email: mbeer@jhu.edu

ABSTRACT

Accurately predicting regulatory sequences and enhancers in entire genomes is an important but difficult problem, especially in large vertebrate genomes. With the advent of ChIP-seq technology, experimental detection of genome-wide EP300/CREBBP bound regions provides a powerful platform to develop predictive tools for regulatory sequences and to study their sequence properties. Here, we develop a support vector machine (SVM) framework which can accurately identify EP300-bound enhancers using only genomic sequence and an unbiased set of general sequence features. Moreover, we find that the predictive sequence features identified by the SVM classifier reveal biologically relevant sequence elements enriched in the enhancers, but we also identify other features that are significantly depleted in enhancers. The predictive sequence features are evolutionarily conserved and spatially clustered, providing further support of their functional significance. Although our SVM is trained on experimental data, we also predict novel enhancers, and show that these putative enhancers are significantly enriched in both ChIP-seq signal and DNaseI hypersensitivity signal in the mouse brain, and are located near relevant genes. Finally, we present results of comparisons between other EP300/CREBBP data sets using our SVM and uncover sequence elements enriched and/or depleted in the different classes of enhancers. Many of these sequence features play a role in specifying tissue-specific or developmental-stage-specific enhancer activity, but our results indicate that some features operate in a general, or tissue independent manner. In addition to providing a high confidence list of enhancer targets for subsequent experimental investigation, these results contribute to our understanding of the general sequence structure of vertebrate enhancers.

INTRODUCTION

Enhancers are gene regulatory sequences that can control transcriptional activities at a distance, independent of their position and orientation with respect to affected genes (Banerji 1981). Enhancer activity is modulated by interactions between sequence specific DNA binding proteins and sequence elements in the enhancer. Since individual transcription factor binding sites (TFBSs) can be relatively short and degenerate, TFBSs tend to be clustered to achieve precise temporal and developmental specificity (Kadonaga 2004). Factors bound to these sequences often interact with common coactivators, which in turn recruit the basal transcription machinery (Blackwood and Kadonaga 1998; Carter et al. 2002).

Identifying the sequence elements and the combinatorial rules that determine enhancer function is necessary to fully understand how enhancers direct the spatial and temporal regulation of gene expression. Experimentally identified enhancers with similar functions can be a good starting point for in-depth study of the underlying rules encoded in the regulatory DNA sequence. However, the systematic functional identification of such enhancers has been limited due to the fact that they are often distant from the genes they regulate, requiring the interrogation of large amounts of potential regulatory sequence. Most investigations make use of two complementary approaches to detect putative regulatory regions: *comparative genomics*, which identifies enhancers by their sequence conservation across related species; and *functional genomics*, which identifies enhancers by the common binding of transcriptionally associated factors or marks (reviewed in Noonan and McCallion 2010).

Comparative genomics is based on the generally accepted hypothesis that functionally important regulatory sequences are under purifying selection. As a result, conserved noncoding sequences (CNS) are natural candidates for putative enhancers. Early studies used CNSs to detect putative enhancers and test their activity in zebrafish or mouse reporter assays (Woolfe et al. 2004; Pennacchio et al. 2006; Visel et al. 2008). Although these conservation based approaches achieve some success, limitations also exist. The function and spatio-temporal specificity of CNSs cannot be determined by conservation alone, and therefore requires additional experimentation. More importantly, several studies have shown that noncoding sequences that apparently lack conservation (as assessed by sequence alignment) may still contain functional regulatory elements (Fisher et al. 2006; ENCODE Project Consortium 2007; McGaughey et al. 2008).

Functional genomics is an experimentally driven approach that utilizes recently developed techniques of microarray hybridization or massively parallel sequencing in combination with chromatin immunoprecipitation (ChIP) on specific transcription factors (Robertson et al. 2007; Johnson et al. 2007), chromatin signatures (Heintzman et al. 2007; Heintzman et al. 2009), or coactivators (Visel et al. 2009; Kim et al. 2010). Specifically, some chromatin signatures or coactivator association (such as mono-methylation of lysine 4 of histone H3 and acetylation of lysine 27 of histone H3, and binding by coactivators EP300/CREBBP) are predictive markers of enhancer activity (Heintzman et al. 2007; Heintzman et al. 2009). The transcriptional coactivators EP300 (also known as P300) and CREBBP (also known as CBP) have proven to be useful for enhancer identification because of their general roles as co-factors in mammalian transcription. Through highly conserved protein-protein interactions, EP300/CREBBP are hypothesized to operate as coactivators in at least three ways: as a direct bridge between sequence specific transcription factors (TFs) and RNA Polymerase II, as an indirect bridge between sequence specific TFs and other co-activators which recruit RNAPol II, or by modifying chromatin structure via intrinsic acetyl-transferase activity (Chan and La Thangue 2001). Several studies have reported genome-wide mapping of EP300/CREBBP bound enhancers in different contexts, for example, tissue-specific activity in dissected mouse tissue (Visel et al. 2009), and environment dependent activity in neurons (Kim et al. 2010). Visel et al. validated that 90% of the EP300 enhancers tested recapitulated the expected spatial and temporal activity *in vivo* in a transgenic mouse enhancer assay. Functionally identified EP300-bound regions thus provide a robust starting point for further investigation of

enhancers and their sequence properties.

In principle, a complete understanding of enhancer mechanism would include a description of specific internal sequence features and how they contribute to enhancer function. Previous studies that have attempted to predict enhancers from sequence have typically used sequence conservation, co-localization of previously characterized TFBSs (from databases such as TRANSFAC (Matys et al. 2003) or JASPAR (Bryne et al. 2008)), or a combination of the two. Many of these existing approaches were assessed by (Su et al. 2010), who found that some were successful in identifying enhancers in *Drosophila*, but that few generalized to mammalian systems. The most successful method in mammalian enhancer prediction used a combination of conservation and low order Markov models of sequence features (Elnitski et al. 2003; King et al. 2005). In more recent work, (Leung and Eisen 2009) used word frequency profile similarity between pairs of sequences to detect novel enhancers, but training on small numbers of enhancers can be susceptible to noise. Another notable recent computational approach uses combinations of known TFBSs and *de novo* position weight matrices (PWMs) to detect enhancers (Narlikar et al. 2010).

In this paper we present a discriminative computational framework to detect enhancers from DNA sequence alone that does not rely on conservation or known TF binding specificities. We use a support vector machine (SVM) to differentiate enhancers from non-functional regions, using DNA sequence elements as features. SVMs (Boser et al. 1992; Vapnik 1995) have been successfully applied in many biological contexts (reviewed in Schölkopf et al. 2004; Ben-Hur et al. 2008): cancer tissue classification (Furey et al. 2000); protein domain classification (Leslie et al. 2002; Karchin et al. 2002; Leslie et al. 2004); splice site prediction (Rätsch et al. 2005; Sonnenburg et al. 2007); and nucleosome positioning (Peckham et al. 2007). In our case, because of the potentially diverse mechanisms which direct EP300 and CREBBP binding, we use a complete set of DNA sequence features to capture combinations of binding sites active in different tissues and times of development. To study these distinct modes of regulation, we investigate EP300/CREBBP binding in mouse embryos (Visel et al. 2009), activated cultured neurons (Kim et al. 2010), and embryonic stem (ES) cells (Chen et al. 2008). Our analysis will initially focus on Visel's data set, where several thousands of EP300-bound DNA elements were collected by ChIP-seq in dissected mouse embryo forebrain, midbrain, and limb. We evaluate our method by predicting enhancers vs. random sequence and *between* EP300/CREBBP ChIP-seq data sets. These comparisons reveal a diversity of predictive sequence features, both within and across data sets. Table S1 provides an outline of the analyses performed in this paper.

We show that sequence features in the experimentally identified enhancer set are sufficient to accurately discriminate enhancers from random genomic regions. We also show that the most predictive sequence elements are related to biologically relevant transcription factor binding sites. Notably, our method also finds that some sequence elements are significantly *absent* in the enhancers (those with large negative SVM weights). For example, we find that binding sites for the zinc finger E-box binding homeobox (ZEB) transcription factor family is depleted in the forebrain enhancers, consistent with its biological role as a transcriptional repressor (Vandewalle et al. 2008). In addition, we provide evidence that enriched sequence elements are positionally constrained within the enhancers, and that they are more evolutionarily conserved than less predictive elements in the enhancers, reflecting the combinatorial structure of tissue-specific enhancers.

We further apply our SVM method to predict putative enhancers in both the mouse genome and the human genome from DNA sequence alone. Many of these novel enhancers overlap with regions enriched in EP300 ChIP-seq reads, exhibit greatly increased hypersensitivity to DNaseI in the mouse brain, and are proximal to biologically relevant genes. All of these assessments exclude the original EP300 training set enhancers from the analysis. The successful identification of tissue-specific DNaseI hypersensitive sites provides powerful independent evidence for the validity of our approach.

RESULTS

Enhancers can be accurately predicted from DNA sequence

Our primary concern in this paper is to identify which sequence features are specific to enhancers, and to investigate the degree to which we can identify functional enhancer regions in a mammalian genome using only DNA sequence features in these regions. We initially focus on recent genome-wide experiments that identified EP300 binding sites by ChIP-seq (Visel et al. 2009) in three different tissues (forebrain, midbrain, and limb) at embryonic day 11.5 in mice. Cross-linking in dissected tissue at a particular time point during development can identify tissue-specific enhancers, even when the developmental regulators that mediate EP300 binding are unknown. While EP300 ChIP may not detect all the enhancers active under these conditions, we initially analyze this dataset to identify sequence features responsible for EP300 binding in these tissues.

To model DNA sequence features, we use a support vector machine (SVM) framework. In brief, an SVM finds a decision boundary that maximally distinguishes two sets of data, here a positive (enhancer) and negative (random genomic) sequence set. The basic approach is outlined in Figure 1A and full details can be found in Methods. Weights, w_i, determine the contribution of each feature to this boundary. Once the set of sequence features, x_i, is specified, the weights are optimized to maximize the separation between the two classes. We use as sequence features the full set of *k*-mers of varying length (3-10). While other authors have successfully used databases of experimentally characterized TFBSs as sequence features (Gotea et al. 2010), because the binding specificity of many transcription factors (TFs) has yet to be determined, we prefer *k*-mers (oligomers of length *k*) because they are an unbiased, general, and complete set of sequence features. An advantage of this framework is that the SVM can be subsequently used to scan the genome for novel enhancers not in the original training set. The results of scanning a well-studied region near *Dlx1/2* is shown in Figure 1B, and detects novel and experimentally confirmed enhancers, as discussed in detail below.

To evaluate classification performance, we use a 5-fold cross validation method. Initially, the data set to be classified is randomly partitioned into five subsets. One subset is then reserved as a test data set, and the SVM weights are trained on sequences in the remaining four subsets. The SVM is then used to predict the reserved test data set to assess its accuracy. This process is repeated five times so that every sequence element is classified in one test set. Because there is a trade-off between specificity (the accuracy of positively classified enhancers) and sensitivity (the fraction of positive enhancers detected), we measure the quality of the classifier by calculating the area under the ROC curve (auROC), as shown for several cases in Figure 2. We ultimately average the five test set auROCs to give a summary statistic of the SVM performance; these five test sets generate the error bars in Figure 2.

To test sensitivity to various assumptions in our SVM construction, we repeated these cross-validation experiments on each tissue-specific enhancer set using SVM classifiers with different types of kernels: Spectrum kernels (Leslie et al. 2002), Mismatch spectrum kernels (Leslie et al. 2004) and Gaussian kernels. The Gaussian kernel and Spectrum kernel vary the functional form by which features contribute to the overall decision boundary, while the mismatch spectrum kernel retains the linear contribution of the features, but uses a different set of features by allowing a certain number of base pair mismatches to a given *k*-mer (see Methods). In addition, we tested a commonly used alternative approach, the Naïve Bayes classifier, which learns the parameters for each feature independently (the SVM learns parameters for all features at the same time). Despite this assumption of independence, the Naïve Bayes classifier has performed very well on a broad range of machine learning applications.

Our main result, perhaps surprising, is that many SVMs can successfully distinguish enhancers from random genomic sequences with auROC > 0.9, regardless of: the types of kernels, the types of tissues, and the length of the *k*-mers (Figure 2 and Figure S1A). In general, larger *k*-mers achieved superior

performance (Figure S1A), but predictive power begins to decrease when *k* is greater than 6 because of overfitting (the feature vector becomes sparse). On the other hand, Naïve Bayes classifiers are significantly less accurate in discriminating enhancers from random genomic sequences (auROC < 0.79), indicating that the assumption of conditional independence between *k*-mers in the Naïve Bayes model impairs its performance. Figure 2A,B, and C show summaries of comparison between ROC curves of SVM (solid) and Naïve Bayes (dotted). Because of its robust performance (auROC=0.94) and ease of interpretation, we adopt the 6-mer spectrum kernel as our standard model for the remainder of the paper.

Besides distinguishing individual enhancer sets from random genomic sequences, we next tested whether our SVM method could also distinguish between enhancers in different tissues (forebrain, midbrain, limb). Since some enhancers are active in two or more tissues, these overlapping regions were removed from both sets before analysis. With the full set of 6-mers, forebrain and midbrain enhancers can be discriminated from limb enhancers with a reasonable auROC of 0.84 ~ 0.86. However the SVM failed to successfully discriminate forebrain and midbrain enhancers (Figure 2D). This indicates that the compositions of TFBSs enriched in forebrain and midbrain enhancers may be similar to each other, but are sufficiently different from those in limb specific enhancers to permit classification. Significant overlap between the forebrain and midbrain enhancer sets in the original data set supports this interpretation (48.7% of midbrain enhancers are also in the forebrain set).

When comparing against random genomic sequence, we have freedom to choose the size of the negative sequence set. The genomic ratio of enhancers to non-enhancer sequence is very large (we estimate that enhancers comprise 1-2% of the genome in a given cell-type), and ideally we would compare alternative prediction methods using a very large negative set. However, some of the computational methods we compared could not handle such large amounts of sequence due to memory constraints. To compare between datasets, we used the same ratio between positives and negatives. To test the scaling with negative set size, we used three negative sets (roughly balanced, 1x; 50x larger; and 100x larger than the positive enhancer set). Although auROC is a standard metric, when the positive and negative sets are unbalanced, the precision-recall (P-R) curve is a more reliable measure of performance than the ROC curve. Precision is the ratio of true positives to predicted positives, and recall is identical to the true positive rate in the ROC curve. The P-R curves can be quantified by the area under the precision-recall curve (auPRC), or average precision. For the classification of EP300 fb, lb, and mb enhancers from genomic sequence, auROC is unaffected by the size of the negative set (Figure 2E), but auPRC drops (Figure 2F) as *n* becomes large and the high scoring tail of the negative sequences becomes competitive with the true positive sequences. However, the trends of auROC and auPRC are usually consistent. Comparison of auROC and auPRC for the negative set size scaling for all positive datasets is shown in Figure S3.

Most predictive sequence elements are known transcription factor binding sites

We next investigated which subsets of sequence features that allowed the SVM to successfully discriminate enhancers from random sequence. The SVM discriminant function is defined as the sum of weighted frequencies of *k*-mers in the case of the *k*-spectrum kernel, and the classification is determined by the sign of the discriminant function (see Methods). Therefore, *k*-mers with large positive and negative SVM weights indicate predictive sequence features: *k*-mers with large positive weights are sequence features specific to enhancer sequences and *k*-mers with large negative weights are sequences that are present in random genomic sequence but depleted in enhancers. We conducted the SVM classification again, using only the subset of *k*-mers with largest positive and negative SVM weights (Figure S1). The SVM using fifty 6-mers with the largest positive weights and another fifty 6-mers with the largest negative weights achieves auROC of 0.90 for the forebrain enhancer data set. This demonstrates that the largest weight *k*-mers predict enhancers with similar accuracy, although the auROC does decrease somewhat compared to the result with all *k*-mers (Figure 2A,B,C). Interestingly, the most frequently observed *k*-mers do not always have the largest SVM weights or vice versa. We find only a weak correlation

between SVM weights and *k*-mer frequencies (Figure S4). The most predictive single *k*-mer (auROC=0.65) is AGCTGC, which is present in 60% of the true positive forebrain enhancers, but it is also present in 34% of the negative genomic regions. By combining many *k*-mers, the full SVM and the SVM with 100 top *k*-mers achieve greater accuracy than single *k*-mers. The SVMs outperformance of the Naïve-Bayes classifier, which assumes feature independence, indicates that these features contribute cooperatively.

Significantly, many of the most predictive *k*-mers, (those with the largest positive weights) are recognizable as binding sites for TFs known to be involved in embryonic nervous system development. We systematically scored each of the predictive *k*-mers with PWMs for known motifs available in public databases (JASPAR (Bryne et al. 2008), TRANSFAC (Matys et al. 2003), and UniPROBE (Newburger and Bulyk 2009)) using the TOMTOM package (Gupta et al. 2007). Because the databases contain many PWMs from families of TFs with similar specificity, many PWMs often score highly for a given *k*-mer, so we report for each *k*-mer the family of matched TFs with *q*-value <0.1 (Storey and Tibshirani 2003), and list representative high scoring TFs within that family. This mapped known TFBS to 85% of the most predictive *k*-mers, while only 24% of all *k*-mers match a known TFBS (Binomial test *p*-value=1.5e-08). Table 1A shows the fifteen 6-mers with the largest positive SVM weights. The full lists of SVM weights used in our analysis are provided in Supplementary Material. The elements that positively contribute to EP300 binding include many *k*-mers with TAAT or ATTA cores, which are bound by the homeodomain family (Berger et al. 2008). Several homeodomain proteins have restricted expression in the embryonic mouse forebrain, and are required for proper forebrain development, such as *Otx* and *Dlx* (Bulfone et al. 1993; Matsuo et al. 1995; Zerucha et al. 2000). Other predictive factors include the members of the basic helix-loop-helix (bHLH) family, which bind variations of E-box elements (CANNTG). Some bHLH factors are known to be crucial regulators of neural and cortical development (Lee 1997; Bertrand et al. 2002; Ross et al. 2003), and are also known to interact with the coactivator EP300/CREBBP (Chan and La Thangue 2001).

One of the distinguishing features of our approach is its ability to detect binding sites that are significantly *absent* or depleted in EP300 enhancers. The presence of *k*-mers with large negative weights in a sequence significantly decreases the likelihood that that sequence will be classified as an enhancer. Biologically, the presence of these binding sites would interfere with the operation of the enhancer in a specific tissue. We consistently observe that ZEB1-related *k*-mers have the largest negative weights in forebrain enhancers (Table 1B). For example, the ZEB1 binding *k*-mer CAGGTA is present in 29% of the negative sequences, but only 18% of the forebrain enhancer sequences. Also known as AREB6, ZEB1 (zinc finger E-box binding homeobox 1) is a member of the ZEB family of transcription factors, which play crucial roles in epithelial-mesenchymal transitions (EMT) in development and in tumor metastasis by repressing transcription of several epithelial genes including E-cadherin (Vandewalle et al. 2008). Although ZEB family members can work as both activators and repressors, their depletion in EP300-bound regions implies that ZEB1 binding can disrupt EP300 activation.

Although some negative weight *k*-mers are predictive (*e.g.* ZEB1), on average the positive weights in Table 1A are more predictive than the negative weights (Table 1B) for all datasets. The absolute values of most negative weight *k*-mers are significantly less than those of the positive weight *k*-mers, as shown in Figure 3 (discussed below), where each *k*-mer weight is plotted along the vertical axis. The asymmetry in SVM weights indicates that the predictive features are primarily identifying *k*-mers that are enriched in the enhancers rather than *k*-mers that are enriched in random genomic sequence (or equivalently, depleted in enhancers).

Predictive sequence elements are evolutionarily conserved and positionally constrained within enhancers

In their previous analysis, Visel *et al*. showed that most EP300-bound regions are enriched in evolutionarily constrained non-coding regions (Visel et al. 2009). However, not all sequences in the EP300-bound regions (average length 750-800bp) are conserved, rather, several more localized peaks of conservation (10-100bp) within the EP300-bound regions are observed in most cases. These peaks of localized conservation probably identify the smaller functional regions within a more extended enhancer. We hypothesized that if the predictive *k*-mers reflect actual TFBSs, they would tend to be preferentially located within these evolutionarily conserved localized regions. To test this systematically, we measured the degree to which individual *k*-mers were present in conserved regions by averaging the PhastCons conservation score (Siepel et al. 2005) over each instance of the *k*-mer (see Methods), and examined its correlation with SVM weight. Figure 3 shows that *k*-mers with large positive SVM weights are significantly more conserved than average. All but one (CCCCTC) of the 6-mers with large positive SVM weights (three or more standard deviations above the mean) have large conservation scores (at least one and a half standard deviation above the mean conservation score). While the most predictive *k*-mers are significantly more conserved, moderate correlation between the PhastCons conservation scores and the SVM weights for all *k*-mers is also observed (Pearson correlation coefficient = 0.35). This evidence supports the idea that the predictive sequence features are more evolutionarily conserved than the less predictive regions within the enhancers.

Since conservation are found in narrow peaks within the enhancers, it follows that there might be additional positional constraints between the predictive elements. Mechanistically, these constraints are most likely indicative of a cooperative mechanism, either involving TF-TF interactions or spatially constrained activity of individual factors. Spatial constraints between TFBSs have been observed frequently in yeast (Beer and Tavazoie 2004). In Figure 4, we compare the distribution of minimum pairwise distances between the ten most predictive sequence elements in the forebrain enhancers (6-mers with the largest positive weights) to their distribution in the null sequences. The forebrain pairwise distance distribution is shifted to lower distances (they are closer to each other) compared to null sequences. To measure the statistical significance of this difference, we calculated the pairwise distance distribution for these 6-mers in 100 different negative sets. The standard deviations of these 100 negative sets are shown as dashed lines in Figure 4, and the forebrain distribution often deviates from the null distribution by several standard deviations, especially for small spacing. We can also measure the difference between the forebrain and null pairwise distance distributions by the two-sample Kolmogorov-Smirnov test, (p-value $< 2.2e-16$), which further demonstrates the significant clustering of predictive sequence elements. More interestingly, if we concentrate on the small spacing end of this distribution (insert in Figure 4), we observe periodic enrichments with characteristic spacing of 10–11bp. The highest peak is around 11bp, almost two times higher than the null distribution. These positional correlations suggest cooperative binding interactions in phase with the 10.5bp DNA helix periodicity, consistent with previous observations (Erives and Levine 2004; Hallikas et al 2006), and local physical interactions between the factors that bind these DNA sequence elements.

Genome-wide SVM predictions identify novel enhancers

To predict additional functional regions that were not determined to be EP300-bound from the ChIP-Seq data, we scanned the entire genome systematically with our SVM. We segmented the mouse genome sequence into 1kbp regions with 0.5kbp overlap, resulting in about 5.2 million overlapping sequence regions. To compare with the 2453 forebrain region “EP300 training set”, we followed Visel and removed centromeric regions, telomeric regions, and regions containing at least 70% repeats, (however, this filter had minimal impact on our predictions). We then scored all these 1kbp regions using the SVM with the k=6 spectrum kernel for forebrain enhancers. An example of the continuous SVM score along the *Dlx1/2* locus

is shown in Figure 1B ("Raw SVM Score"). *Dlx1* and *2* are expressed in the mouse forebrain (Bulfone et al. 1993; Ghanem et al. 2003; Wigle and Eisenstat 2008). Besides the sole EP300 training set element in this region (URE2, labeled "EP300 ChiPseq" in Figure 1B), two other enhancers within this locus have been experimentally validated ("Known Enhancers", labeled i12a and i12b) (Ghanem et al. 2003). These enhancers (i12a and i12b) were detected by our SVM, but were not in the EP300 training set because their raw sequence read density was not above the stringent threshold used in (Visel et al 2009). Comparing the "Raw EP300 ChIPseq" track to our "Raw SVM score" in Figure 1B shows striking correlation: most of our predicted high scoring SVM regions have raw EP300 ChIP-seq signal significantly above background, but did not have sufficient read density to be included in the EP300 training set. To support this anecdotal evidence, we evaluated the genome-wide correlation between our SVM predicted regions and EP300 read density. In Figure S5 we plot the EP300 ChIP-seq read density as a function of distance from the center of each of the top 1% SVM scoring regions. We find significant enrichment of EP300 ChIP-seq signal around the SVM predicted regions, indicating that many of these predicted loci are indeed bound to some extent by EP300, but fall somewhat below the read threshold used to determine the EP300 training set. Figure S6 shows the correlation between SVM score and EP300 reads in all genomic 1kbp regions, showing again that there is a significant population of high scoring SVM regions enriched in EP300 signal but not in the EP300 training set.

To define a high confidence set of enhancer predictions, we chose an appropriate cutoff for the SVM score using more realistic large negative training set sizes (50x and 100x negative sequences), covering about 6-12% of the non-repetitive genome. We can estimate our false discovery rate (the expected fraction of predicted positives which are false positives, FP/(FP+TP)) from the P-R curves in Figure 2F. The precision is weakly dependent on negative set size when n is large, due to the fact that the positive and negative histograms of SVM scores have similar shape for larger negative set sizes, as shown in Figure S7. To trade off precision and recall, we choose a cutoff which corresponds to 50% recall, which at 1x is an SVM score of 1.0. For the large negative sets, precision is about 50% when recall is 50%, and we therefore estimate our false discovery rate to be about 50%. In other words, at this cutoff (SVM>1.0), on the training set, we capture 50% of the EP300 training set regions, and an equal number of negative regions.

In what follows we will be comparing the properties of our SVM predicted enhacer regions (SVM>1.0), the EP300 training set regions, and non-enhancer genomic regions (SVM<1.0). These three sets are all distinct, i.e. each genomic 1kbp region can only belong in one class. Any 1kbp region which overlaps a training set region by as little as 1bp is excluded from the SVM sets and included in the EP300 training set. We will show that the EP300 training set and SVM predicted regions have similar properties, much different than the non-enhancer regions.

At an SVM score threshold of 1.0, we predict 33,232 1kbp regions in the genome (outside of the EP300 training set), or 26,920 enhancers after merging overlapping regions, and we expect about 13,460 of these to be true enhancers. This threshold appears to be a good tradeoff between detecting many biologically significant enhancers with an acceptable false discovery rate. The full lists of SVM scores for these regions are included as Supplementary Material. We also established the robustness of these top SVM scoring regions by training separate SVMs with independent random null sequence sets as the negative class. There is extensive overlap between the top scoring regions using these different SVMs (Table S2), and the correlation of individual SVM scores between two different SVMs is high (Pearson correlation coefficient=91.5%), as shown in Figure S8. That the SVM classifier identifies many more sequence regions than the EP300 training set may be due to several factors: 1) As discussed above these predicted regions may be false positive enhancers, 2) They may be true positive enhancers that were undetected in the ChIP experiments because of an overly stringent cutoff for defining the EP300 training set, 3) They may be true positive enhancers that are not EP300-bound in this tissue at the developmental stage of the experiment, but may be EP300-bound in other tissues or times, 4) They may be true positive enhancers that operate independently of EP300, but share some similar sequence features. All but the first possibility are

potentially biologically interesting.

To assess the validity these genome-wide predictions with independent experimentation, we quantified the DNaseI hypersensitivity of the high scoring forebrain SVM regions with experiments in embryonic mouse whole brain provided by the mouse ENCODE project (data available from http://genome.ucsc.edu/ENCODE/, Stamatoyannopoulos 2011), using methods described in (John et al. 2011). DNaseI hypersensitivity measurements detect open or accessible chromatin, including promoters and enhancers, independent of EP300 binding. Although these DNaseI experiments are not strictly specific to forebrain and were three days later in development, enrichment in brain hypersensitivity strongly corroborates our predictions as tissue specific enhancers. In Figure 5, we split the predicted 1kbp regions from the EP300 fb trained SVM into 4 classes (SVM<0.5 red, 0.5<SVM<1.0 grey, 1.0<SVM<1.5 cyan, and SVM>1.5 blue) and one EP300 training set class (EP300-bound regions, green). We plot the distributions of average intensity of DNaseI hypersensitivity of the different SVM scoring classes in Figure 5A, which shows a dramatic increase in DNaseI signal in E14.5 brain only for high scoring SVM regions. There is no enrichment of DNaseI signal for the same regions in other tissues, for example adult kidney is shown in Figure 5B as a negative control. Because the DNaseI hypersensitive regions include promoters and other open regions, the converse is not true, i.e. while almost all high scoring SVM regions have a high DNaseI signal, not all high signal DNaseI regions have a high SVM score (data not shown). With this understanding,we can evaluate the precision and specificity with which our SVM detects DNaseI sensitive enhancers. Because the SVM score and DNaseI signals are continuous, we consider DNaseI signal > 10 to be positive (open chromatin), and DNaseI < 2 to be negative (not open) for purposes of quantification, consistent with the distributions in Figure 5A and B. Then, regions with DNaseI > 10 and SVM >1.0 are true positive predictions, and DNaseI < 2 and SVM >1.0 regions are false positive predictions. Table 2 shows the number of 1kbp genomic regions in each class. The precision is TP/(TP+FP), or the accuracy of the predicted positives. The sensitivity is 1-FPR (false positive rate), or the fraction of negatives that we predict to be positive. As shown in Table 2, SVM>1.0 predictions have a 56.3% precision, and more stringent SVM>1.5 predictions have a 74.5% precision. These results are consistent with our above estimate that 50% of our novel predictions are true enhancers functioning in mouse brain.

To further support the biological significance of these novel SVM predicted enhancers, we examined their proximity to forebrain expressed genes. Microarray experiments (Visel et al. 2009) identified 885 (495) genes over-expressed (under-expressed) in the forebrain at E11.5. We examined the intergenic distance between the EP300 training set regions and the transcription start site (TSS) of the nearest over-expressed genes. We also found the distance between our SVM predicted enhancer regions and the over-expressed genes. All regions overlapping a training set region were omitted from the set of predictions. As shown in Figure 6, both the EP300 training set and our predicted enhancer regions are significantly enriched near (within 10kbp) of the TSS of a forebrain over-expressed gene. Notably, the SVM predicted regions with the more stringent SVM cutoff score (SVM>2.0) are even more enriched within 10kbp of the over-expressed genes than the EP300 training set, further evidence that the SVM is capturing functional regions with spatial and temporal specificity. In comparison, randomly chosen genomic regions show no such enrichment. While the EP300 training set is not enriched near forebrain *under*-expressed genes, our SVM predicted regions are significantly enriched within 10kbps of forebrain under-expressed genes (Figure 6). What is a potential role of these predicted regions near under-expressed genes? Because the EP300 bound regions are not enriched near the under-expressed genes, it is unlikely that EP300 is acting as a transcriptional repressor here. It seems more likely that the SVM is predicting enhancers that are bound by EP300 in other tissues or at other times in development. These enhancers could activate the neighboring genes relative to their expression level at E11.5 in the forebrain, which would appear indistinguishable from forebrain repression. This hypothesis is supported by the fact that several of the under-expressed genes with nearby SVM predicted enhancers play roles in nervous system development, including many Hox genes known to function in A-P axis patterning.

SVM also predicts human enhancers

We next assessed the ability of our SVM to predict human enhancers. We found human orthologous regions (hg18) of the mouse EP300 training set with the liftOver utility from the UCSC genome browser (Karolchik et al. 2008). With 70% or greater identity, 2205 of the 2453 forebrain enhancers were successfully mapped onto the human genome. We discarded 13 mapped sequences longer than 3kbps. We then trained SVMs to discriminate this positive human training set from an equal number of human random sequences generated by our null model, and achieved reasonably high auROC=0.87 (Figure S9). We also tested more stringent orthology cutoffs (requiring 90% and 95% identity instead of 70%), and found that the overall performance was very similar (Figure S9). Thus an SVM trained on human sequence homologous to the mouse EP300 training set sequences is able to predict test set enhancers with only slightly reduced accuracy relative to mouse.

In addition, we predicted human enhancer regions with a SVM trained on the mouse dataset, which does not require sequence alignment to identify orthologous regions. This approach might be valuable in situations where it is difficult or impossible to obtain similar datasets in each species. It also provides further information about the conservation of predictive *k*-mers between the two species. We first compared these two raw SVM scores (one trained on human homologous set, the other on mouse dataset) on the human genome around *Otx2*, observing very similar SVM score patterns. Moreover, an experimentally verified enhancer (Kurokawa et al. 2004) is captured by both SVMs (Figure S10). We then systematically analyzed the entire genome to assess how many top SVM scoring regions overlap each other (Table S3). Although the overlaps are not as significant as scores using only different negative sets (Table S2), a large fraction of top SVM scoring regions are still shared between the two SVMs, so to a large degree, an SVM trained on mouse can be used to successfully predict human enhancers. This result is in general agreement with *in vivo* experimental results (Wilson et al. 2008) where human DNA transplanted into mice was shown to bind mouse TFs (HNF1A, HNF4A, HNF6) in a pattern virtually indistinguishable from their binding patterns in human, indicating that variations in genomic TF binding between human and mouse are due to local DNA sequence differences, not due to evolutionary divergence of individual TF binding specificities between the two species.

Comparison between different EP300/CREBBP ChIP-seq data sets reveals sequence elements important for pluripotency

The success of our SVMs in predicting EP300 binding in mouse embryonic brain and limb motivated a comparison with other EP300/CREBBP ChIP-seq data sets. We first looked at the overlap between Visel's *in vivo* data set (EP300 forebrain, midbrain, and limb) and two other data sets: CREBBP bound regions in activated cultured mouse cortical neurons (Kim et al. 2010), and EP300-bound regions in cultured mouse embryonic stem cells (Chen et al. 2008). We will refer to these as "CREBBP neuron" and "EP300 ES" in the following discussion. We were interested in these datasets because they share similar ChIP-seq methodology, because it would help us address the overlap between activation mediated by the close homologs EP300 and CREBBP, and to address differences in EP300 binding in different tissues and cell populations. CREBBP neuron enhancers only overlap significantly with EP300 forebrain enhancers (not midbrain or limb, Table S4A). EP300 ES enhancers do not significantly overlap with any other set (fb, mb, lb, or CREBBP neuron, Table S4B). This indicates that EP300 mediated embryonic neuronal development is linked to CREBBP mediated neural activity dependent transcription via extensively shared common regulatory regions. We indeed observe that several predictive *k*-mers with large positive weights, such as homeodomain binding sites (TAAT core) and bHLH domain binding sites (E-box, CANNTG), are shared between the two data sets (Table 1A and Table S5A), which further indicates common modes of regulation.

Figure 2G shows ROC curves discriminating CREBBP neurons (auROC=0.93) and EP300 ES

(auROC=0.77) from random genomic sequences. The lower EP300 ES auROC is partly due to the relatively smaller number of regions bound in the EP300 ES positive set. Also, the EP300 ES data set contains a larger fraction of repeat sequences, indicating that this data set may be less specific for functional EP300 binding. Nonetheless, SVMs still can extract informative *k*-mers from this data set and can largely discriminate the EP300 ES set from random genomic sequences. Alternatively, instead of comparing vs. random genomic sequence, we can also successfully classify these sets (EP300 forebrain, CREBBP neuron, EP300 ES) against each other, as show in Figure 2H. It is interesting to note that EP300 forebrain can be discriminated from CREBBP neuron with high auROC, even though they share many regions and have some common predictive *k*-mers (homeodomain, SOX, bHLH) when classified against random sequence (Table 1A and Table S5A). However, when classified against each other, we observe that the predictive *k*-mers specific for EP300 forebrain remain homeodomain, SOX, and bHLH, but the *k*-mers predictive for CREBBP neurons become Nuclear Factor I (NFI), Activator protein 1 (AP1), and Cyclic AMP-responsive element-binding protein (CREB) binding sites (Table S7). Therefore, homeodomain, SOX, and bHLH binding sites may play more prominent roles in neural developmental processes than in neural activity dependent transcription.

We also assessed the biological significance of the predictive *k*-mers in these new data sets. We find that most of the predictive *k*-mers can be related to known TFBSs (Table S5 and Table S6), and that many of the identified TFBSs are involved in signaling pathways known to function in the relevant experimental conditions. For the CREBBP neuron data set, AP1 related 6-mers, GACTCA and TGACTC, the first and third largest weights respectively (Table S5A), are the target of heterodimers of the regulators Fos and Jun, which play critical roles in neural activity dependent transcription regulation (Flavell and Greenberg 2008). CREB, which directly interacts with CREBBP, is also essential for the activation of several genes in response to neural stimulation, and its binding site is ranked fourth in Table S5A (Flavell and Greenberg 2008; Kim et al. 2010). Kim et al. noted that two other transcription factors, neuronal PAS domain-containing protein 4 (NPAS4) and serum response factor (SRF) as well as CREB, strongly co-localize with CREBBP binding regions. NPAS4 contains a bHLH domain, and its canonical binding sites, E-box elements, are ranked at second and sixth in Table S5A. The SRF binding site is also known as a CArG box, whose consensus sequence is CCWTATAWGG (Bryne et al. 2008). A specific *k*-mer instance of the CArG box is ATATGG, ranked at 17th with w=3.00, just below the top fifteen in Table S5A. Therefore, all well characterized TFBSs known to play a role in neuronal activation are successfully captured by our SVM. Interestingly, we discovered that two additional transcription factor families also score highly in the CREBBP neuron data set: homeodomain and NFI. These families have been discussed little in this context, although it is known that both NFI and homeodomain transcription factors are key regulators of central nervous system development (Mason et al. 2008; Wilson and Koopman, 2002). We found only one relevant example of neural activity dependent expression of a homeobox protein, LMX1B (Demarque and Spitzer 2010). There may be still unknown mechanisms involving NFI and homeodomain proteins in the context of neural activity dependent transcriptional regulation, but broadly speaking, our results indicate significant pleiotropy between neuronal developmental pathways and neural activity dependent signaling pathways.

Comparison of the EP300 ES data to CREBBP neuron and EP300 forebrain can address which binding sites and factors are responsible for maintaining a differentiated or pluripotent state. For the EP300 ES data set, our method identifies factors known to be crucial for maintaining ES identity: we find high scoring binding sites for NANOG-POU5F1(also known as OCT4)-SOX2 SOX-family factors (Table S6A), essentially the same binding sites found in previous studies (Pavesi et al. 2001; Chen et al. 2008). We have used a uniform approach to map *k*-mers to TFBS in the databases, but there is substantial overlap in many TF specificities, and some reported matrices may score higher than the biologically relevant database entry. For instance, in Table S6A the high scoring matrices (SOX17, POU2F1, and POU3F3) appear on the list instead of the relevant (SOX2, POU5F1, and NANOG) which have nearly identical binding sites. SOX2,

POU5F1, and NANOG bind a combination of the SOX2 (CATTGT) and POU5F1 (ATGCAAAT) consensus sites (Chen et al. 2008), and the 6-mer subsequences within the combined binding site (CATTGTYATGCAAAT) have high SVM weights. Figure S11 shows how large weight *k*-mers tile across this extended known binding site. In addition, we also find positive weight binding sites for ESRRB and STAT3, which are known to be frequently located nearby the NANOG-POU5F1-SOX2 clusters assessed by ChIP-seq analysis (Chen et al. 2008). More interestingly, we find that many of the positive weight EP300 ES *k*-mers (ESRRB, RORA1/2, PPARG) are among the largest negative weights in CREBBP neuron (Table S6B), indicating that binding sites for factors responsible for maintaining pluripotency are significantly absent from neuronal enhancers (CREBBP neuron), as would be expected given the developmental maturity of neurons.

SVM can predict other ChIP-seq data sets

Until this point we have applied our SVM method to classify and detect EP300/CREBBP-bound enhancers, but this approach is equally applicable to any dataset which may be framed as a sequence classification: e.g. ChIP-seq, ChIP-chip, or DNaseI hypersensitivity datasets. In these situations the SVM can be used to identify primary binding sites in regions identified by transcription factor ChIP experiments, and may also identify binding sites for secondary factors co-localized with the ChIPed TF, or binding sites significantly depleted in the functionally occupied regions. We note that popular *de novo* motif finding methods such as AlignACE (Hughes et al. 2000) or MEME (Bailey and Elkan 1994) have limited success when applied to data sets of this size. When run on the forebrain enhancer data set, AlignACE (when it converged) failed to report any meaningful motifs. While Chen et al. (Chen et al. 2008) did successfully identify Sox2, Oct4, and Nanog binding sites in the EP300 ES data with Weeder (Pavesi et al. 2001), the EP300 ES data set was the smallest and least diverse of the data sets we analyzed.

To directly assess the ability of our SVM to predict binding of individual transcription factors, we analyzed ChIP-seq results on the TF ZNF263. We chose ZNF263, a 9-finger C2H2 zinc finger which is predicted to have a binding site of ~24bp, to assess how well *k*-mers can represent extended degenerate binding sites. We used ChIP-seq data on ZNF263 in a K562b cell line (Frietze et al. 2010) which identified 1418 strongly bound regions. Predicting against a 50x random negative set yielded auROC=0.938 and auPRC=0.51 (Figure S12B,D). Many of the largest weight *k*-mers are subsequences within the large PWM found by *de novo* motif finding tools applied to this data set (Frietze et al. 2010), and the SVM is combining *k*-mers which tile across the binding site to achieve high predictive accuracy. The *k*-mer GAGCAC also received a large weight. This indicates that our approach should have significant predictive value for a wide range of binding data.

Comparison to alternative approaches

As an alternative to *k*-mers, we also tried using known PWMs as features in an SVM. We used 811 PWMs from existing databases of known TF specificities (JASPAR (Bryne et al. 2008), TRANSFAC (Matys et al. 2003), and UniPROBE (Newburger and Bulyk 2009)). When using these features, we used the highest PWM scores in each sequence for each matrix as the feature vector. This 811 PWM SVM was able to achieve auROC=0.87 for forebrain enhancers (compared to auROC=0.93 for *k*-mers), somewhat less predictive than our *k*-mer approach (Figure S12A), against a 50x random negative set. However, this translates into a significantly lower auPRC=0.22 (compared to auPRC=0.43 for *k*-mers) (Figure S12B). The optimal combined weighting of the known PWMs and 6-mers features (2080+811 total features) gives marginal improvement (auROC=0.93 and auPRC=0.49) over 6-mers alone. We also applied the 811 PWM SVM to the ZNF263 dataset, which achieved auROC=0.83 (compared to auROC=0.94 for *k*-mers), reflecting the fact that accurate PWMs for ZNF263 were absent from the databases (Figure S12B,D). Again this seemingly small change in auROC corresponds to a large drop in auPRC=0.14, compared to auPRC=0.51 for *k*-mers. This demonstrates that using sequence features from an unbiased and complete

set can be more valuable than using an incomplete set of more accurate features (PWMs). Using the set of known TF PWMs is less predictive than our *k*-mer SVM, but a more complete set of PWMs might perform better. Combining the predictive *k*-mers into a more general PWM, via a method similar to POIMs (Sonnenburg et al. 2008), might allow clearer identification of informative sequence features from within the *k*-mer SVM, but would not affect predictive performance.

We also compare our approach to alternative kernel methods. We applied the weighted degree kernel with shifts (WDS) (Rätsch et al. 2005) to the CREBBP neuron data set (as WDS requires input sequences of equal length), and found auROC=0.83, compared to auROC=.93 for our *k*-mer SVM. A notable SVM based approach which incorporates positional information between general *k*-mer features (KIRMES) has been recently described (Schultheiss et al. 2009; Schultheiss 2010). We applied this package to the forebrain EP300 dataset and found auROC=0.90. In the current implementation of KIRMES, *k*-mers are selected by their relative frequency in the positive set, and it is likely that further optimization would make this approach comparable to our *k*-mer SVM result. Additionally, the periodic spatial distribution in Figure 4 suggests that a model based on difference in angle (similar to Hallikas et al. 2006) would be more appropriate than the Gaussian spatial dependence used in KIRMES. Another approach to predict promoters (Megraw et al. 2009) used PWMs and l1-logistic regression. We found little difference between logistic regression and SVM: using our *k*-mer feature vectors in l1-logistic regression yielded auROC=0.92 on the EP300 forebrain dataset, using publicly available software (Koh et al. 2007).

DISCUSSION

In this study, we have shown that a support vector machine (SVM) can accurately predict regulatory sequences without any prior knowledge about transcription factor binding sites (TFBSs), using only general genomic sequence information. While the ROC and P-R curves demonstrate that the SVM is able to identify enhancers based on their sequence features, the biological relevance of the predicted enhancers is further supported by: 1) Most of the predictive sequence features identified by our methods are binding sites of previously characterized TFBSs known to play a role in the relevant context. 2) The enriched predictive sequence features are much more evolutionarily conserved within the enhancers than the less predictive sequence features, which suggests that the predictive features are under selection and comprise the functional subset of the larger enhancer regions. 3) These sequence features are significantly more spatially clustered in the enhancers than would be expected by chance, also a well-known characteristic of functional binding sites. 4) Genomic regions with high forebrain SVM scores are strongly enriched in DNaseI hypersensitivity signals in mouse brain, but not in other tissues. 5) The predicted enhancers frequently overlap with regions of enhanced ChIP-seq signals, but are somewhat below the signal cutoff necessary to be included in the original EP300 training set. 6) These novel predicted enhancers are preferentially positioned near biologically relevant genes, and many have been experimentally verified in other studies, which further supports their biological relevance and functional roles.

When scanning the whole genome to predict putative enhancers, we predict that 50% of our 26920 non-overlapping enhancers with forebrain SVM scores above 1.0 are true positives. This is a conservative estimate of our ability to detect novel enhancers, since when scanning the genome we have scored 1kb arbitrarily delimited chunks of sequence: more accurate predictions might be possible by varying the endpoints of the predicted regions. Nevertheless, this genome-wide scan discovers thousands of novel predicted enhancers that were not in the original experimental training set. We have shown that we can predict human enhancers based on these mouse enhancer experiments by measuring the overlap between human enhancers predicted by an SVM trained on the mouse sequence, and comparing these predictions to an SVM trained on human sequence orthologous to the mouse enhancer sequences. Finally, by comparing between other EP300/CREBBP ChIP-seq data sets, we find sequence features that are able to differentiate between enhancers that operate in different tissues or at different developmental stages. Some of these sequence features are enriched in enhancers in one specific tissue or state, but other predictive elements are

notably depleted in some classes of enhancers.

It is perhaps surprising that such a simple description of sequence features (*k*-mer frequencies) is able to classify enhancers and ChIP-seq data so well. The SVM is apparently combining *k*-mer features in a sufficiently flexible way to reflect combinations of binding sites and/or sequence signals which modulate chromatin accessibility. Developing an optimal sequence feature vector remains an area for future work, however, our results showing that the SVM is more accurate than Naïve-Bayes suggests that successful prediction requires the ability to combine features without evaluating them independently.

Several features of our results suggest ways that our method could be improved to make more accurate predictions. It is likely that incorporating positional constraints between the features would improve the accuracy of the predictions, consistent with our observation of non-random spatial distributions between predictive features in the SVM. Kernel approaches have been developed which incorporate positional information, but most have been developed in the context of positional constraints relative to a single preferred genomic location or anchor point. In application to other problems, positional information relative to a transcription start site (Sonnenburg et al. 2006b), to a splice site (Rätsch et al. 2005; Sonnenburg et al. 2006; Sonnenburg et al. 2007), or to a translational start site (Meinicke et al. 2004) has been implemented in SVM contexts. Positional preference relative to a mean anchor point has been incorporated in a *de novo* motif discovery method developed by (Keilwagen et al. 2011). However, the aforementioned methods are not strictly appropriate to the biological problem of enhancer detection, because enhancers have no such preferred fixed location, and the relevant positional constraints are *between* sequence features within the enhancer. Many approaches have modeled clusters of known binding sites (reviewed in Su et al. 2010), but have limited application to mammalian enhancer prediction.

Although we have provided evidence that our SVM predicted regions are likely functional, to what degree we are predicting these enhancers accurately based on sequence features which are tissue-specific? Alternatively, we could be detecting sequence features which are general to larger classes of enhancers. These common features could allow access, stabilize, or could be recognized by generic components of the enhanceosome (Thanos and Maniatis 1995; Maniatis et al. 1998), whose activity could be modulated by tissue-specific factors, much as Pol II operates generally. Ultimately this should be determined by individual experiments, but we here address this problem computationally by investigating overlaps between forebrain and limb-specific predicted regions, which we then compare with the overlaps between EP300-enriched regions in forebrain and limb. For this comparison, we independently determined EP300-enriched regions from the raw data set using the same threshold criteria as the previous study (Visel et al. 2009) except that we have used fixed-length 1kbp regions, rather than the ChIP-seq determined peak regions. With a 1% false discovery rate (FDR), we obtained 3390 EP300-enriched regions of forebrain and 2607 regions of limb. Visel's EP300-bound regions are highly tissue-specific; there are only 243 regions (7-9%) shared by the two sets. For the SVM predictions, a significantly larger fraction of forebrain predicted regions (6104 out of 39714, 15%) are found in 34% of the limb predicted regions (18027). This suggests that our SVMs learn features that are generally enriched in enhancers, in addition to tissue-specific sequence features. As a result, two SVMs trained on entirely different data sets can predict common regions that have general enhancer function. Moreover, the 6104 regions predicted by both limb and forebrain SVMs overlap with small EP300 peaks that are somewhat below the conservative threshold (FDR<0.01); almost 50% have peak in at least one tissue. This observation further supports our hypothesis that SVM predicted regions are likely to be functional. A further complication is that individual tissues consist of heterogeneous populations of cell types, and enhancers predicted in distinct tissues may only be active in subsets of cell types. A detailed analysis of which sequence features impart tissue specificity and which are general is suggested as a focus for future investigations.

METHODS

Data Sets

As positive data sets, we initially used the genome-wide *in vivo* EP300 binding sites identified by ChIP-seq (Visel et al. 2009), composed of three different sets of tissue-specific enhancers (forebrain, midbrain, and limb) of embryonic day 11.5 mouse embryos. 2453, 561 and 2105 sites were reported, respectively, and we directly use the entire sequences without modification. We also analyzed two other data sets (Kim et al. 2010; Chen et al. 2008). Chen *et al*. reported 524 EP300 binding sites in mouse embryonic stem cells, and Kim *et al.* reported ~12000 neural activity dependent CREBBP binding sites in stimulated cultured mouse cortical neurons. Since both CREBBP data sets report only peaks of the ChIP-seq signals, we extended 100bp (for Figure 2G) or 400bp (for Figure 2H) in both directions from these peaks to obtain sequences for further analysis.

We generated negative sequence sets to match the distribution of sequence length and repeat element fraction of the corresponding positive sets (Figure S2). Repeat fractions were calculated using the repeat masked sequence data from the UCSC genome browser (Karolchik et al. 2008). We selected random genomic sequences from the mouse genome according to the following rejection sampling algorithm:

1. Sample a length l from the enhancer length distribution.

2. Sample a sequence of the length l, randomly from the genome.

3. Let x be the repeat fraction of the sampled sequence. Sample $Y \sim Bernoulli(\alpha p(x)/q(x))$, where $p(x)$ is the probability that x occurs in the enhancers, $q(x)$ is the probability that x occurs in the genomic sequence, α is the constant so that the maximum of $p(x)/q(x)$ equals 1.

4. Accept the sequence if Y=1, reject otherwise.

5. Repeat 1-4 until the desired number of sequences are sampled.

All positive and negative sequence data sets used for our analysis are available at http://www.beerlab.org/p300enhancer. We used the following negative set sizes: EP300 fb: n=4000, 2453(1x), 122650(50x), 245300(100x); EP300 mb: n=4000, 561(1x), 28050(50x), 56100(100x); EP300 lb: n=4000, 2105(1x), 105250(50x), 210500(100x); EP300 fb human n=2192(1x); EP300 ES n= 524(1x), 5240(10x), 26200(50x), 52400(100x); CREBBP neuron n=11847(1x), 592350(50x), 1184700(100x); ZNF263 n=1418(1x), 70900(50x), 141800(100x).

Support Vector Machine

An SVM (Boser et al. 1992; Vapnik 1995) finds a decision boundary that separates the positive and negative training data. This decision boundary is a hyperplane which maximizes the margin between the two sets in the feature vector space. We have N labeled vectors $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, N$, where $\mathbf{x}_i \in R^n$ and $y_i \in \{+1, -1\}$ is the class label. For the linear case, the decision boundary is found by minimizing $\|\mathbf{w}\|^2$ such that $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1, i = 1, \cdots, N$. In practice, the optimal solution is found by maximizing the dual form: $\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i \alpha_i y_j \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ over α_i with the constraints, $\alpha_i \geq 0$, and $\sum_{i=1}^{N} \alpha_i y_i = 0$ (Joachims 1999; Sonnenburg et al. 2006). The SVM weight vector $\mathbf{w}$ can be constructed

from the α_i using $\mathbf{w} = \sum_{i=1}^{N} y_i \alpha_i \mathbf{x}_i$. The SVM discriminant function, or "SVM score", $f_{\mathrm{SVM}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \sum_{i=1}^{N} y_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}) + b$, represents the distance of any vector **x** from the decision boundary, and determines the predicted label of the vector **x**.

The inner product $(\mathbf{x}_i \cdot \mathbf{x}_j)$ is a measure of the similarity of any two data points *i* and *j* in the feature space. The generality of the SVM arises from the fact that this term may be replaced by a more general measure of similarity, a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. Different kernels refer to different methods of measuring similarity. A very simple and general measure of sequence similarity is the *k*-spectrum kernel (Leslie et al. 2002), which describes the similarity of *k*-mer frequencies of two sequences. We have found that this kernel produces our best results, is easy to interpret, and can easily represent a combination of TF binding sites. To implement the *k*-spectrum kernel, we generate a *k*-mer count vector for the full set of distinct *k*-mers, for each sequence. Then we normalize the count vector so that ||**x**||=1, to reduce the effect of the variable length of different enhancers. We loosely refer to this normed vector as the "*k*-mer frequency vector." The kernel function is then just the inner product between two normalized frequency vectors. To reflect the fact that TFs bind double stranded DNA , the spectrum kernel function is slightly modified to account for both orientations. Instead of counting only an exact *k*-mer, its reverse complement is also counted, and then redundant *k*-mers are removed. For example, only one of AATGCT and AGCATT appears on the list of distinct *k*-mers. For 6-mers, there are 2080 distinct features after removing reverse complements; for 7-mers there are 8192. This modification was applied to all kernel functions. The only difference between the *k*-spectrum kernel and the (*k*,*m*)-mismatch kernel is that the mismatch kernel allows *m* mismatches when counting *k*-mers (Leslie et al. 2004), reflecting the fact that some TFs bind degenerate sites. The Gaussian kernel uses the same feature vectors as the *k*-spectrum kernel, but uses a non-linear similarity measure via the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$. Our implementation utilizes the Shogun machine learning toolbox (Sonnenburg et al. 2006) and SVM light (Joachims 1999). The full lists of SVM weights are provided in Supplementary Material, and python scripts are available from our website http://www.beerlab.org/p300enhancer.

AUTHOR CONTRIBUTIONS

DL conceived of the study as a final project in RK's Foundations of Computational Biology course at JHU, DL and MB carried out the analysis, DL and MB wrote the paper, and all authors read and approved the manuscript.

ACKNOWLEDGEMENTS

We thank Donavan Cheng, Mahmoud Ghandi, Rahul Karnik, Changhee Lee, and Andy McCallion for useful discussions and helpful comments. We also appreciate detailed suggestions from the anonymous reviewers which significantly improved the manuscript. We thank J. Stamatoyannopoulos and his lab for generating and allowing pre-publication access to the mouse ENCODE DNaseI data. M. Beer was supported by the Searle Scholars Program and in part by NS062972 (NIH). R. Karchin was supported in part by NSF DBI-1845275.

FIGURE LEGENDS

Figure 1 – Overview

Outline of our methodology. (A) *k*-mer frequencies are calculated for each of the EP300-bound and negative genomic training sequence. These feature vectors ($\mathbf{x}_1,\ldots,\mathbf{x}_n$) are used to find SVM weights $\mathbf{w}$ which most accurately separate the positive (enhancer) and negative (genomic) training sets. (B) These weights are used to predict genome-wide enhancers (light green), based on their SVM score (brown-positive and blue-negative). A well studied region around *Dlx1* and *Dlx2* is shown here, both known to be expressed in forebrain. While the predicted enhancers often overlap the training EP300 set (blue), novel enhancers are also predicted, and often identify previously experimentally verified enhancers (red), absent from the EP300 training set. The predicted enhancers also preferentially occur in conserved non-exonic regions (dark green) and regions enriched in EP300 signal (dark blue).

Figure 2 – ROC curves

Classification results on each tissue-specific enhancer set. (A) Classification of forebrain enhancers vs. random genomic sequences. (B) Classification of midbrain enhancers vs. random genomic sequences. (C) Classification of limb enhancers vs. random genomic sequences. Each graph in A,B and C compares an SVM trained on the full set of 6-mers (solid), the top 100 selected 6-mers (dashed), and an alternative Naïve Bayes classifier (dotted). Each curve is an average of 5 cross-fold validations on a reserved test set; error bars denote one standard deviation over the 5 cross-fold validation sets. Numbers in parenthesis indicate the area under each ROC curve (auROC) for overall comparison. Both the full SVM and SVM with selected features perform very well, and significantly better than Naïve Bayes. Individually, each tissue-specific set can be accurately discriminated from non-enhancer genomic sequences. (D) Classification of specific tissues vs. other tissues. Forebrain (fb) and midbrain (mb) can be accurately discriminated from limb (lb), but not from each other (fb vs mb), indicating common or overlapping modes of regulation. (E) Classification ROC curves for forebrain enhancers vs. random genomic sequences for larger negative set sizes. (F) Precision-Recall curves for forebrain enhancers vs. random sequences corresponding to the ROC curves and negative sets in (E), numbers in parenthesis are auPRC. (G) Classification of EP300 forebrain enhancers, neuronal stimulus dependent enhancers (CREBBP neuron), and mouse embryonic stem cell enhancers (EP300 ES) vs. random genomic sequence. Although the embryonic stem cell dataset is somewhat less accurately classified, our SVMs successfully discriminate EP300 or CREBBP bound regions from random sequences. (H) Classification of EP300 fb, CREBBP neuron, and EP300 ES datasets vs. each other is also robust.

Figure 3 – SVM weights vs. conservation scores

Predictive SVM sequence features are more conserved. Scatter plot between SVM weights and conservation scores (phastCons scores) for 6-mers in forebrain enhancers. Two well known TFBS, TAAT cores (red rectangles) and E-box elements (blue triangles) are highlighted. Three standard deviations above the mean (corresponding to p-value of ~0.001) is denoted for each axis independently. The sequence of all 6-mers beyond three standard deviations above the mean are displayed.

Figure 4 – Distributions of minimum pairwise distances

Predictive SVM sequence features are spatially clustered. Distributions of minimum pairwise distances between the most predictive sequence features in forebrain enhancers vs. random genomic

sequences. Ten 6-mers with the largest positive SVM weights (Table 1) are used. To measure the significance of these differences, we generated 100 distinct full negative genomic sequence sets (using our null model, see Methods). Each negative set has same the length, repeat fraction, and number of sequences as the EP300 forebrain enhancer training set. The predictive elements are significantly clustered in the forebrain enhancers compared with the random genomic sequences (the red distribution is significantly shifted toward smaller minimum distance). At higher resolution (inset), distinct peaks around 11bp, 22bp, etc., are observed, suggesting positioning in phase with the periodicity of the DNA helix. P-values are indicated: * <0.01, ** <0.001, *** <0.0001.

Figure 5 – Distributions of Average Intensity of the DNaseI Hypersensitivity

SVM predicted regions are hypersensitive to DNaseI in the relevant context. To independently confirm our predictions with DNaseI measurements in embryonic mouse brain, we plot the distributions of average intensity of DNaseI hypersensitivity of different forebrain SVM scoring regions. (A) DNaseI Hypersensitivity measured in E14.5 Wholebrain (B) DNaseI Hypersensitivity measured in adult 8 weeks kidney, as a negative control. We observe significant enrichments only in high scoring SVM predicted regions in brain.

Figure 6 – Genome-wide distributions of SVM predicted regions

SVM predicted enhancers are preferentially located near transcript start sites (TSS) of forebrain expressed genes. Here we plot the distribution of the distance between the EP300 and SVM predicted regions and the nearest forebrain expressed gene (as assessed by the microarray experiments of (Visel et al. 2009)). Any region which overlapped a training set region was excluded from the analysis. Both the EP300 (red) and SVM predicted regions are preferentially located within 10kbp of the TSS of a forebrain *over-expressed* gene (above the axis). This is true whether we use a cut-off of SVM>1.5 (green) or a more restrictive SVM>2.0 (blue) to define the enhancer set. As a null set we compare to the average of 100 randomized genomic positions, with 95% confidence interval shown (grey). Interestingly, when we calculate the same distributions for the distance between a EP300 or SVM predicted region and the nearest forebrain *under-expressed* gene (below the axis), only the SVM predicted regions show significant clustering toward the TSS, relative to the randomized control. Although the EP300 data is preferentially identifying activating enhancers in forebrain, the SVM may be detecting common sequence features shared in enhancers which are repressive in forebrain, but are activating in other contexts.

TABLES

Table 1 – Predictive 6-mers of EP300 Forebrain

(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)
AATGAG	CTCATT	3.94	Homeodomain	POU6F1
AATTAG	CTAATT	3.85	Homeodomain	VSX2, PRRX2, EVX2, PDX1, GBX2
AGCTGC	GCAGCT	3.65	HLH	NHLH1, HEN1, ASCL2, REPIN1, TCF3
CAATTA	TAATTG	3.62	Homeodomain	BARHL2, PRRX2, NKX2-5, NKX6-1, BARHL1
CAGCTG	CAGCTG	3.32	HLH	NHLH1, HEN1, REPIN1, ASCL2, MYOD1, TCF3
ACAAAG	CTTTGT	3.29	SOX	SOX4, SOX11, SOX10, HNF4A
TAATTA	TAATTA	3.24	Homeodomain	OTP, PROP1, HOXA, ALX1, LHX3
CAGATG	CATCTG	3.15	HLH	ZFP238, TAL1:TCF3, TAL1:TCF4, TCF3
TAATGA	TCATTA	3.03	Homeodomain	POU6F1, POU4F3, LHX3, HOXC9, NKX6-3
AATTAA	TTAATT	2.94	Homeodomain	LHX3, OTP, PRRX2, PROP1, LHX5
ATTAGC	GCTAAT	2.90	Homeodomain	VSX2, POU3F2, EVX2, PITX3, LHX8
GGCAAC	GTTGCC	2.86	-	-
ACAATG	CATTGT	2.63	SOX	SOX17, SOX9, SOX5, SOX10, SOX30
CATTCA	TGAATG	2.45	SOX	HBP1
AATTAC	GTAATT	2.18	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)
AGGTAG	CTACCT	-1.79	-	-
AAGTCA	TGACTT	-1.89	-	-
AGGTGA	TCACCT	-1.97	Zinc-finger	ZEB1
ACCTGG	CCAGGT	-2.03	Zinc-finger	ZEB1, TCF3
CAGGTA	TACCTG	-2.06	Zinc-finger	ZEB1

Table 2 – Precision and Sensitivity of Detecting DNaseI hypersensitive enhancers.

		True Positives	False Positives	Precision	Sensitivity
		DNaseI > 10	DNaseI < 2	TP/(TP+FP)	1-FP/N
Predicted Positives	SVM>1.5	3892	1330	74.5%	0.9996
	SVM>1.0	11081	8612	56.3%	0.997
Predicted Negatives	SVM<1.0	98590	3086512	3.5%	
		P=109671	N=3095124		

REFERENCES

Bailey T, and Elkan C. 1994. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Banerji J. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299-308.

Beer MA, and Tavazoie S. 2004. Predicting Gene Expression from Sequence. *Cell* **117**: 185-198.

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, and Rätsch G. 2008. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput Biol* **4**: e1000173.

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. 2008. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell* **133**: 1266-1276.

Bertrand N, Castro DS, and Guillemot F. 2002. Proneural genes and the specification of neural cell types. *Nat Rev Neurosci* **3**: 517-530.

Blackwood EM, and Kadonaga JT. 1998. Going the Distance: A Current View of Enhancer Action. *Science* **281**: 60-63.

Boser BE, Guyon IM, and Vapnik VN. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, *COLT '92*, p. 144–152, ACM, New York, NY.

Bryne JC, Valen E, Tang ME, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, and Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucl Acids Res* **36**: D102-106.

Bulfone A, Puelles L, Porteus M, Frohman M, Martin G, and Rubenstein J. 1993. Spatially restricted expression of Dlx-1, Dlx-2 (Tes-1), Gbx-2, and Wnt- 3 in the embryonic day 12.5 mouse forebrain defines potential transverse and longitudinal segmental boundaries. *J Neurosci* **13**: 3155-3172.

Carter D, Chakalova L, Osborne CS, Dai Y, and Fraser P. 2002. Long-range chromatin regulatory interactions in vivo. *Nat Genet* **32**: 623-626.

Chan HM, and La Thangue NB. 2001. P300/CBP proteins: HATs for transcriptional bridges and scaffolds. *J Cell Sci* **114**: 2363-2373.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang

J, et al. 2008. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133**: 1106-1117.

Demarque M, and Spitzer NC. 2010. Activity-Dependent Expression of Lmx1b Regulates Specification of Serotonergic Neurons Modulating Swimming Behavior. *Neuron* **67**: 321-334.

Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, and Chiaromonte F. 2003. Distinguishing Regulatory DNA From Neutral Sites. *Genome Res* **13**: 64 -72.

ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

Erives A, and Levine M. 2004. Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci* **101**: 3851-3856.

Fisher S, Grice EA, Vinton RM, Bessling SL, and McCallion AS. 2006. Conservation of RET Regulatory Function from Human to Zebrafish Without Sequence Similarity. *Science* **312**: 276 -279.

Flavell SW, and Greenberg ME. 2008. Signaling Mechanisms Linking Neuronal Activity to Gene Expression and Plasticity of the Nervous System. *Annu Rev Neurosci* **31**: 563-590.

Frietze S, Lan X, Jin VX, and Farnham PJ. 2010. Genomic Targets of the KRAB and SCAN Domain-containing Zinc Finger Protein 263. *J Biol Chem* **285**: 1393 -1403.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, and Haussler D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**: 906-914.

Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, Park BK, Rubenstein JLR, and Ekker M. 2003. Regulatory Roles of Conserved Intergenic Domains in Vertebrate Dlx Bigene Clusters. *Genome Res* **13**: 533 -543.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, and Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**: 565 -577.

Gupta S, Stamatoyannopoulos JA, Bailey TL, and Noble W. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: R24.

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, and Taipale J. 2006. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. *Cell* **124**: 47-59.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311-318.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108-112.

Hughes JD, Estep PW, Tavazoie S, and Church GM. 2000. Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**: 1205-1214.

Joachims T. 1999. Making large-scale support vector machine learning practical. p. 169–184, MIT Press, Cambridge, MA.

John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL, and Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**: 264-268.

Johnson DS, Mortazavi A, Myers RM, and Wold B. 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**: 1497-1502.

Kadonaga JT. 2004. Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell* **116**: 247-257.

Karchin R, Karplus K, and Haussler D. 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **18**: 147-159.

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.

Keilwagen J, Grau J, Paponov IA, Posch S, Strickert M, and Grosse I. 2011. De-Novo Discovery of Differentially Abundant Transcription Factor Binding Sites Including Their Positional Preference. *PLoS Comput Biol* **7**: e1001070.

Kim T, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182-187.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, and Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* **15**: 1051 -1060.

Koh K, Kim S-J, and Boyd S. 2007. An Interior-Point Method for Large-Scale l1-Regularized Logistic Regression. *J Mach Learn Res* **8**: 1519–1555.

Kurokawa D, Kiyonari H, Nakayama R, Kimura-Yoshida C, Matsuo I, and Aizawa S. 2004. Regulation of Otx2 expression and its functions in mouse forebrain and midbrain. *Development* **131**: 3319-3331.

Lee JE. 1997. Basic helix-loop-helix genes in neural development. *Curr Opin Neurobiol* **7**: 13-20.

Leslie C, Eskin E, Cohen A, Weston J, and Noble WS. 2004. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20**: 467-76.

Leslie C, Eskin E, and Noble WS. 2002. The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput* 564-75.

Leung G, and Eisen MB. 2009. Identifying Cis-Regulatory Sequences by Word Profile Similarity. *PLoS ONE* **4**: e6901.

Maniatis T, Falvo JV, Kim TH, Kim TK, Lin CH, Parekh BS, and Wathelet MG. 1998. Structure and Function of the Interferon-β Enhanceosome. *Cold Spring Harb Symp Quant Biol* **63**: 609 -620.

Mason S, Piper M, Gronostajski RM, and Richards LJ. 2008. Nuclear Factor One Transcription Factors in CNS Development. *Mol Neurobiol* **39**: 10-23.

Matsuo I, Kuratani S, Kimura C, Takeda N, and Aizawa S. 1995. Mouse Otx2 functions in the formation and patterning of rostral head. *Genes Dev* **9**: 2646-2658.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. 2003. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucl Acids Res* **31**: 374-378.

McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, and McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res* **18**: 252 -260.

Megraw M, Pereira F, Jensen ST, Ohler U, and Hatzigeorgiou AG. 2009. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Research* **19**: 644 -656.

Meinicke P, Tech M, Morgenstern B, and Merkl R. 2004. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics* **5**: 169.

Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, Nobrega MA, and Ovcharenko I. 2010. Genome-wide discovery of human heart enhancers. *Genome Res* **20**: 381 -392.

Newburger DE, and Bulyk ML. 2009. UniPROBE: an online database of protein binding

microarray data on protein-DNA interactions. *Nucl Acids Res* **37**: D77-82.

Noonan JP, and McCallion AS. 2010. Genomics of Long-Range Regulatory Elements. *Annu Rev Genom Human Genet* **11**: 1-23.

Pavesi G, Mauri G, and Pesole G. 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17 Suppl 1**: S207-214.

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, and Weng Z. 2007. Nucleosome positioning signals in genomic DNA. *Genome Research* **17**: 1170 -1177.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499-502.

Rätsch G, Sonnenburg S, and Schölkopf B. 2005. RASE: recognition of alternatively spliced exons in C.elegans. *Bioinformatics* **21 Suppl 1**: i369-77.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth* **4**: 651-657.

Ross SE, Greenberg ME, and Stiles CD. 2003. Basic Helix-Loop-Helix Factors in Cortical Development. *Neuron* **39**: 13-25.

Schölkopf B, Tsuda K, and Vert JP. 2004. *Kernel methods in computational biology*. The MIT press, Cambridge, MA.

Schultheiss SJ. 2010. Kernel-Based Identification of Regulatory Modules. In *Computational Biology of Transcription Factor Binding* (ed. I. Ladunga), Vol. 674 of, pp. 213-223, Humana Press, Totowa, NJ.

Schultheiss SJ, Busch W, Lohmann JU, Kohlbacher O, and Rätsch G. 2009. KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. *Bioinformatics* **25**: 2126 -2133.

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.

Sonnenburg S, Rätsch G, Schäfer C, and Schölkopf B. 2006. Large Scale Multiple Kernel Learning. *J Mach Learn Res* **7**: 1531-1565.

Sonnenburg S, Schweikert G, Philips P, Behr J, and Ratsch G. 2007. Accurate splice site prediction using support vector machines. *BMC Bioinformatics* **8**: S7.

Sonnenburg S, Zien A, and Ratsch G. 2006. ARTS: accurate recognition of transcription starts in human. *Bioinformatics* **22**: e472-480.

Stamatoyannopoulos JA, in preparation. 2011.

Storey JD, and Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100**: 9440 -9445.

Su J, Teichmann SA, and Down TA. 2010. Assessing Computational Methods of Cis-Regulatory Module Prediction. *PLoS Comput Biol* **6**: e1001020.

Thanos D, and Maniatis T. 1995. Virus induction of human IFNbeta gene expression requires the assembly of an enhanceosome. *Cell* **83**: 1091-1100.

Vandewalle C, Roy F, and Berx G. 2008. The role of the ZEB family of transcription factors in development and disease. *Cell Mol Life Sci* **66**: 773-787.

Vapnik VN. 1995. *The Nature of Statistical Learning Theory*. Springer, New York, NY.

Visel A, Blow MJ, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**: 854-858.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, and Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**: 158-160.

Wigle JT, and Eisenstat DD. 2008. Homeobox genes in vertebrate forebrain development and disease. *Clin Genet* **73**: 212-226.

Wilson M, and Koopman P. 2002. Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators. *Curr Opin Genetics Dev* **12**: 441-446.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavare S, and Odom DT. 2008. Species-Specific Transcription in Mice Carrying Human Chromosome 21. *Science* **322**: 434-438.

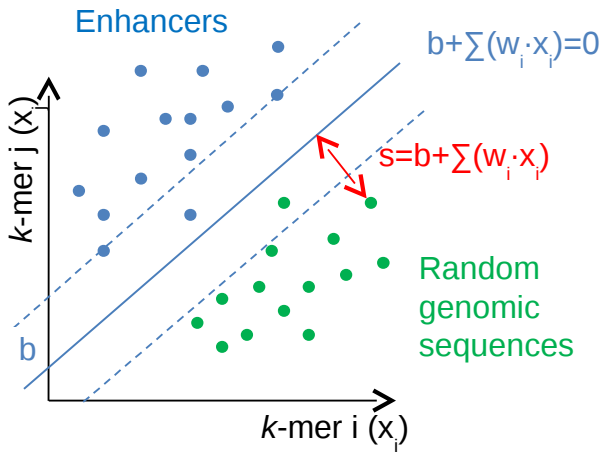
Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2004. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biol* **3**: e7.

Zerucha T, Stühmer T, Hatch G, Park BK, Long Q, Yu G, Gambarotta A, Schultz JR, Rubenstein JLR, and Ekker M. 2000. A Highly Conserved Enhancer in the Dlx5/Dlx6 Intergenic Region is the Site of Cross-Regulatory Interactions between Dlx Genes in the Embryonic Forebrain. *J Neurosci* **20**: 709-721.

A

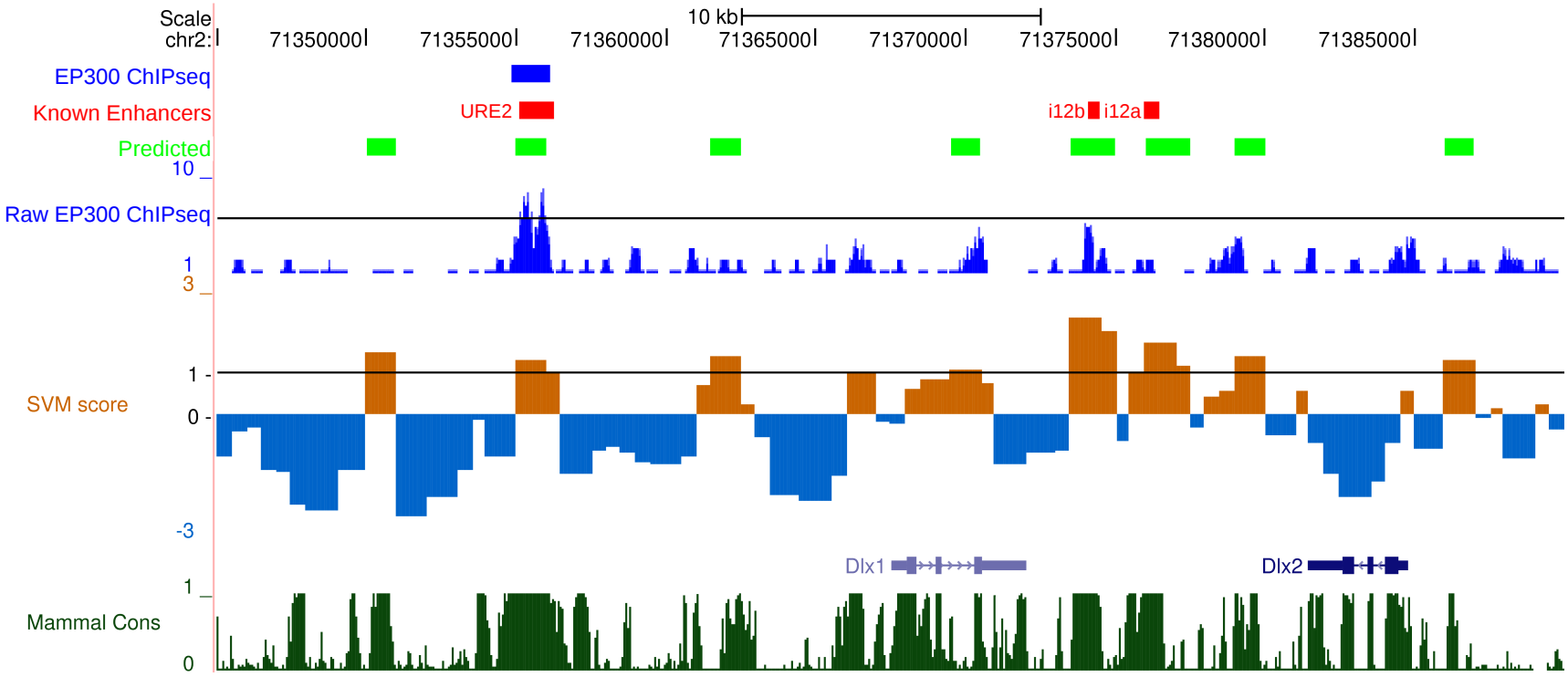
Sequences → *k*-mer frequencies → SVM training → Predictive feature analysis

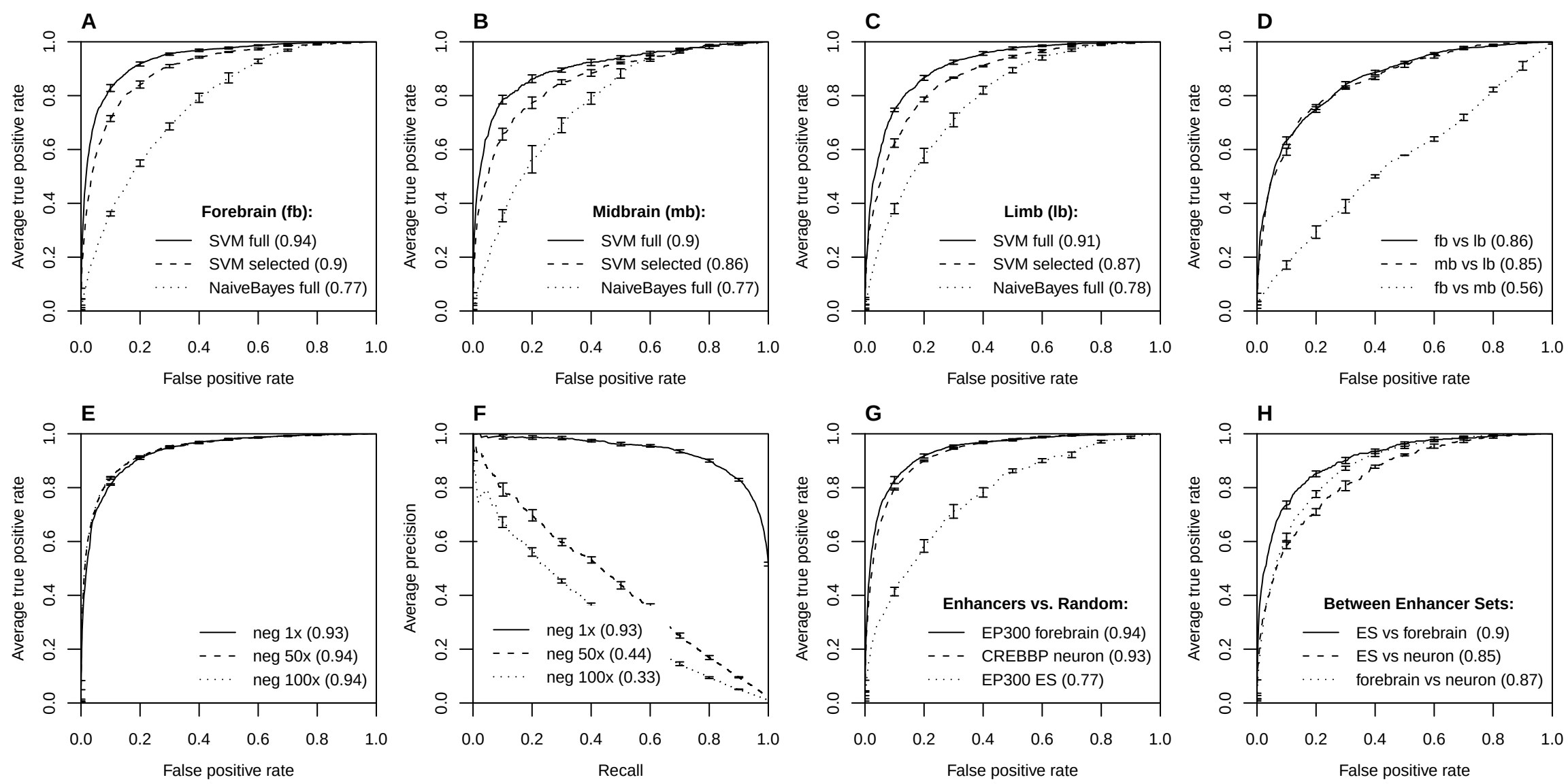
k-mer	counts
5′-AAAAAA-3′ 3′-TTTTTT-5′	x_1
5′-AAAAAC-3′ 3′-TTTTTG-5′	x_2
5′-AAAAAG-3′ 3′-TTTTTC-5′	x_3
...	...
5′-TTTAAA-3′ 3′-AAATTT-5′	x_n

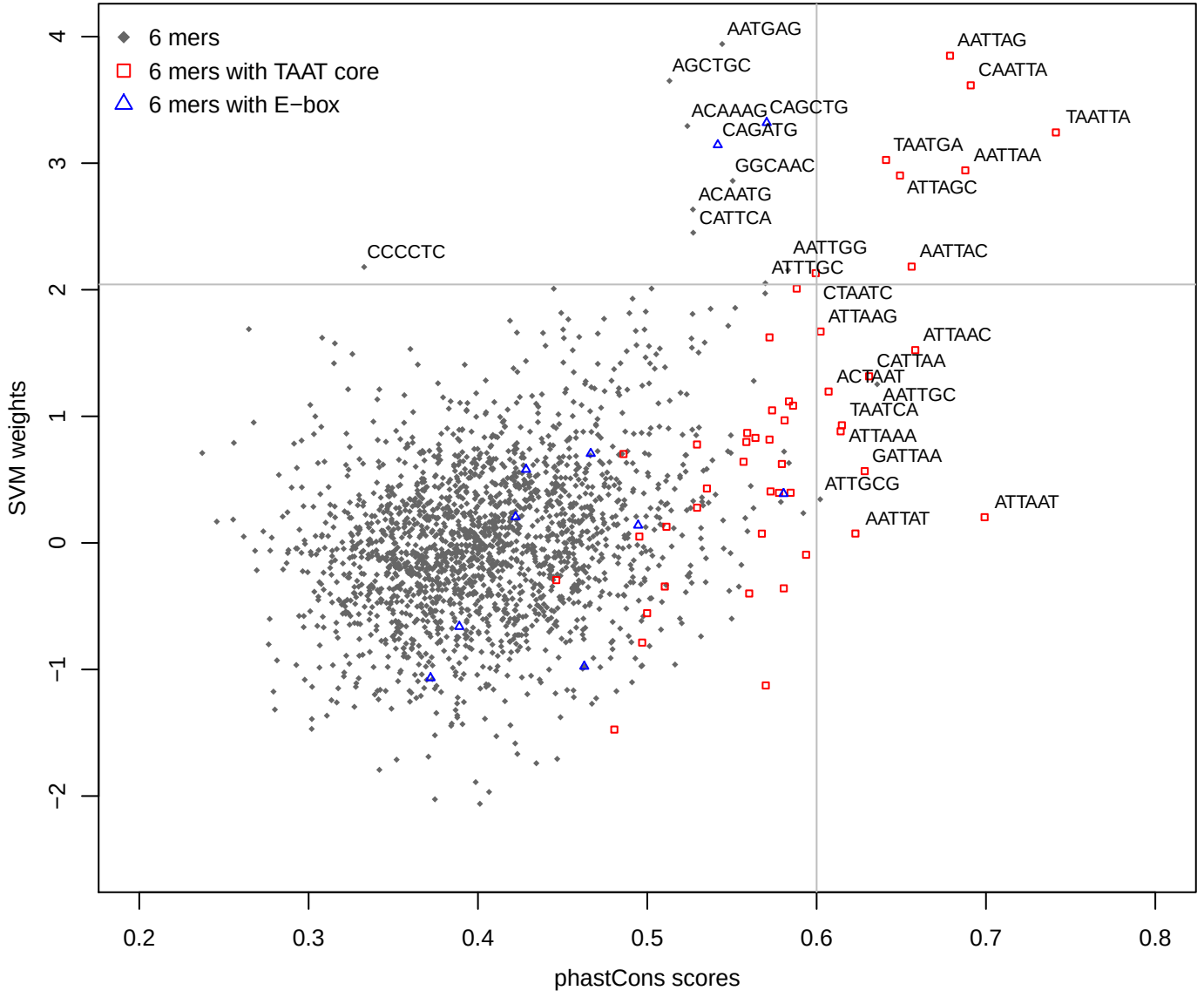


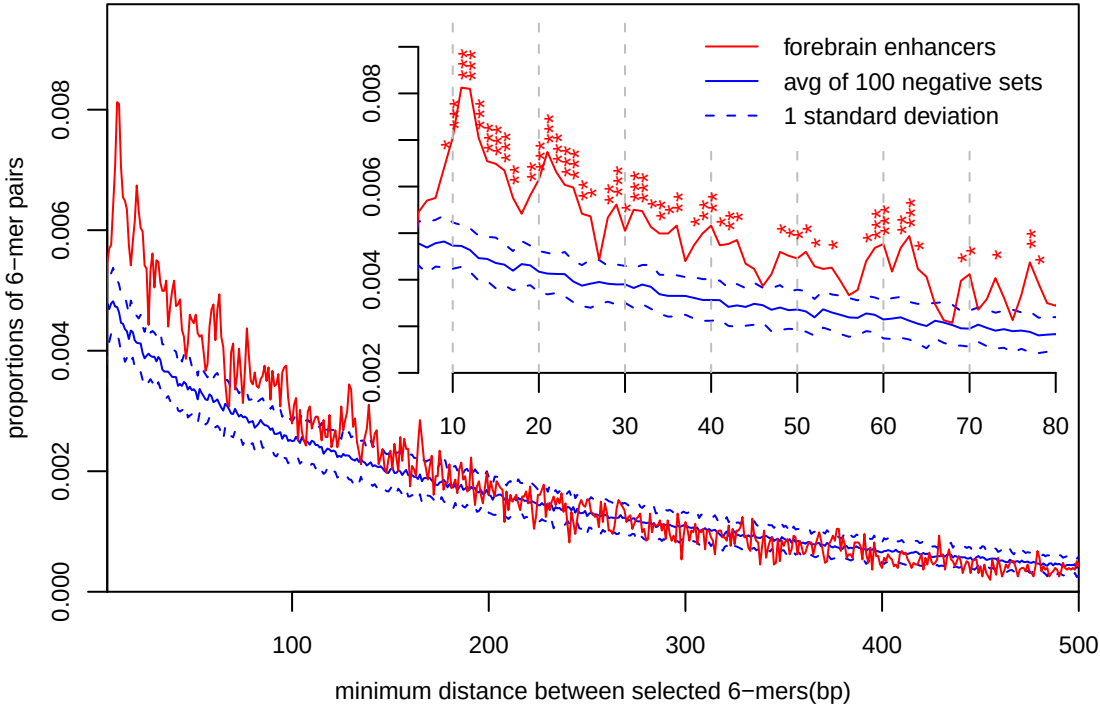
k-mer	weights (w_i)
5′-AATGAG-3′ 3′-TTACTC-5′	+3.94
5′-AATTAG-3′ 3′-TTAATC-5′	+3.84
5′-AGCTGC-3′ 3′-TCGACG-5′	+3.65
...	...
5′-CAGGTA-3′ 3′-GTCCAT-5′	-2.06

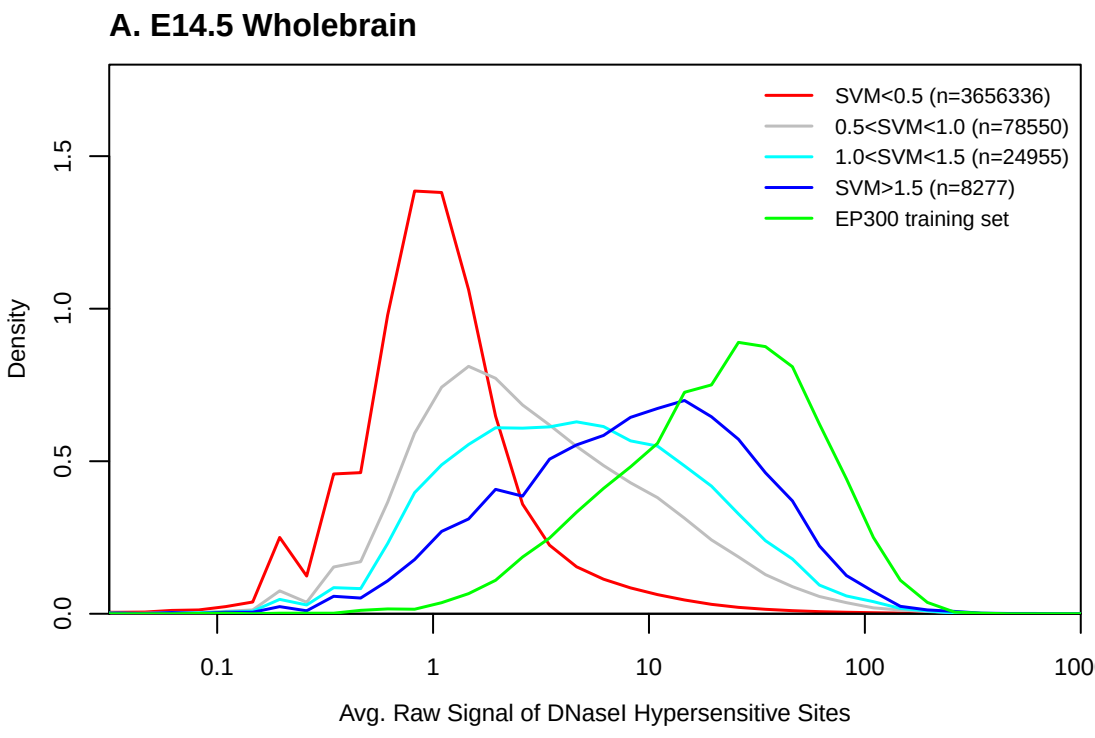
B

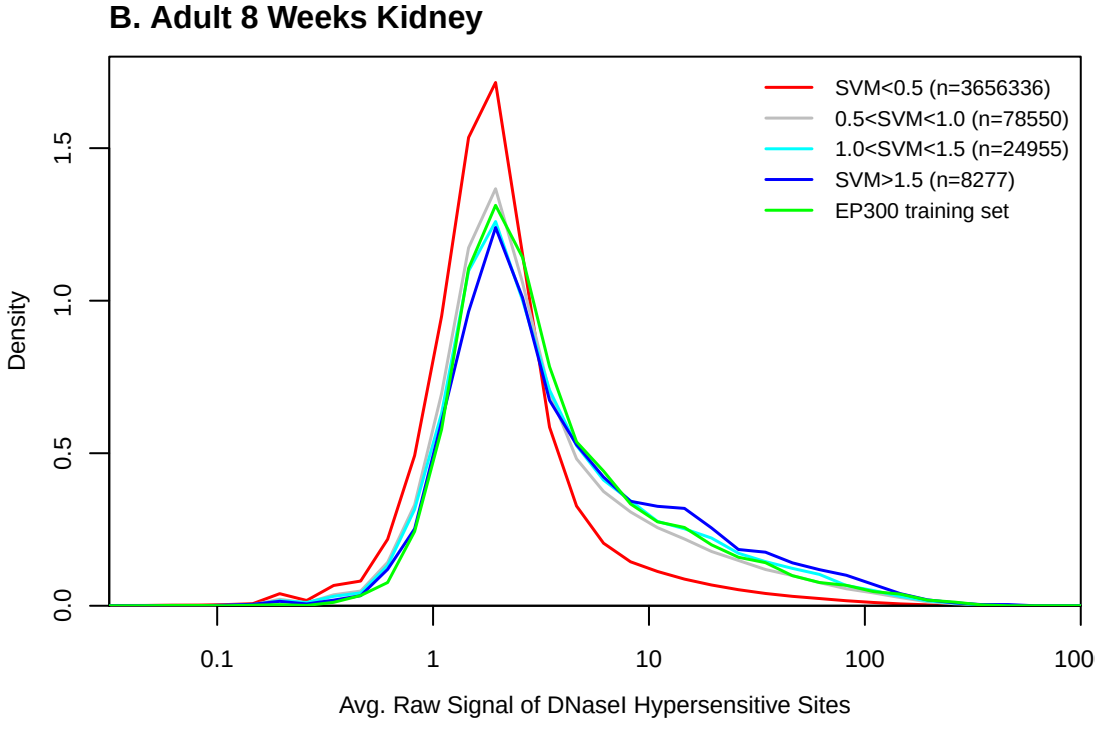


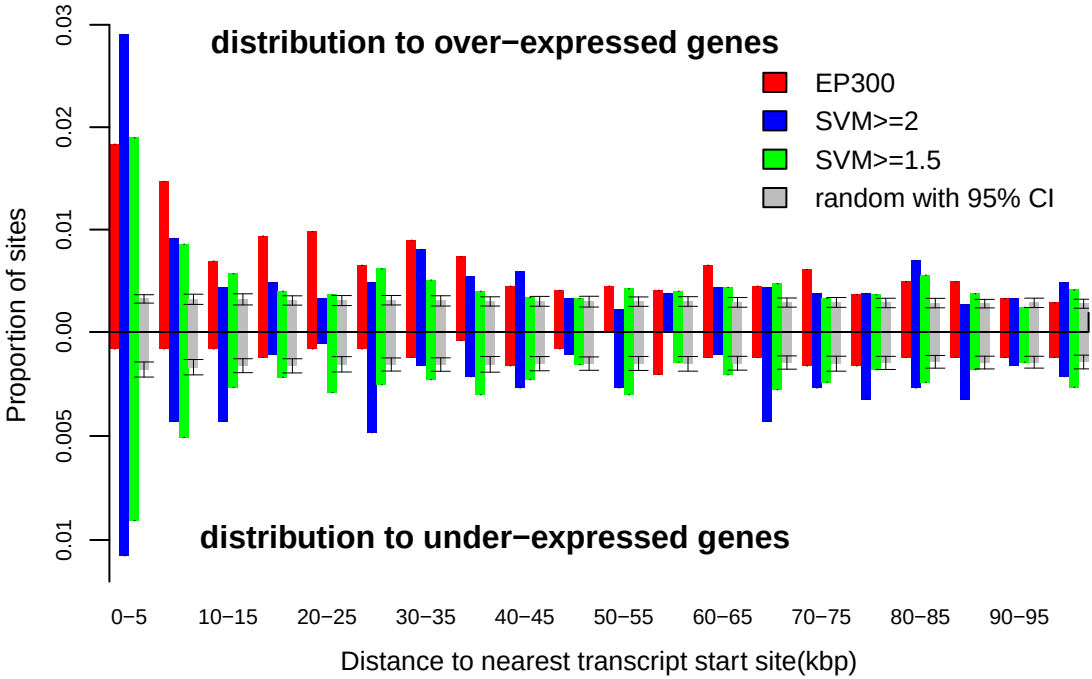

A
Average true positive rate
False positive rate
Forebrain (fb):
SVM full (0.94)
SVM selected (0.9)
NaiveBayes full (0.77)
B
Average true positive rate
False positive rate
Midbrain (mb):
SVM full (0.9)
SVM selected (0.86)
NaiveBayes full (0.77)
C
Average true positive rate
False positive rate
Limb (lb):
SVM full (0.91)
SVM selected (0.87)
NaiveBayes full (0.78)
D
Average true positive rate
False positive rate
fb vs lb (0.86)
mb vs lb (0.85)
fb vs mb (0.56)
E
Average true positive rate
False positive rate
neg 1x (0.93)
neg 50x (0.94)
neg 100x (0.94)
F
Average precision
Recall
neg 1x (0.93)
neg 50x (0.44)
neg 100x (0.33)
G
Average true positive rate
False positive rate
Enhancers vs. Random:
EP300 forebrain (0.94)
CREBBP neuron (0.93)
EP300 ES (0.77)
H
Average true positive rate
False positive rate
Between Enhancer Sets:
ES vs forebrain (0.9)
ES vs neuron (0.85)
forebrain vs neuron (0.87)


6 mers
6 mers with TAAT core
6 mers with E−box
AATGAG
AGCTGC
ACAAAG
CAGCTG
CAGATG
GGCAAC
ACAATG
CATTCA
CCCCTC
AATTAG
CAATTA
TAATTA
TAATGA
AATTAA
ATTAGC
AATTGG
ATTTGC
AATTAC
CTAATC
ATTAAG
ATTAAC
CATTAA
ACTAAT
AATTGC
TAATCA
ATTAAA
GATTAA
ATTGCG
ATTAAT
AATTAT
SVM weights
phastCons scores
4
3
2
1
0
−1
−2
0.2
0.3
0.4
0.5
0.6
0.7
0.8


forebrain enhancers
avg of 100 negative sets
1 standard deviation
proportions of 6−mer pairs
0.000
0.002
0.004
0.006
0.008
100
200
300
400
500
minimum distance between selected 6−mers(bp)
0.002
0.004
0.006
0.008
10
20
30
40
50
60
70
80


A. E14.5 Wholebrain
SVM<0.5 (n=3656336)
0.5<SVM<1.0 (n=78550)
1.0<SVM<1.5 (n=24955)
SVM>1.5 (n=8277)
EP300 training set
Density
0.0
0.5
1.0
1.5
0.1
1
10
100
1000
Avg. Raw Signal of DNaseI Hypersensitive Sites


B. Adult 8 Weeks Kidney
SVM<0.5 (n=3656336)
0.5<SVM<1.0 (n=78550)
1.0<SVM<1.5 (n=24955)
SVM>1.5 (n=8277)
EP300 training set
Density
0.0
0.5
1.0
1.5
0.1
1
10
100
1000
Avg. Raw Signal of DNaseI Hypersensitive Sites


distribution to over−expressed genes
EP300
SVM>=2
SVM>=1.5
random with 95% CI
Proportion of sites
0.03
0.02
0.01
0.00
0.005
0.01
distribution to under−expressed genes
0−5
10−15
20−25
30−35
40−45
50−55
60−65
70−75
80−85
90−95
Distance to nearest transcript start site(kbp)